

Agricultural Statistics

for
M. Sc. Ag. CLASSES

by
S. P. Singh M.Sc.
●
R. P. S. Verma M.Sc.
Department of Statistics
J.V. College, Baraut (Meerut)

Published by
RAMA PUBLISHING HOUSE
BARAUT (MEERUT)

Published by
RAMA PUBLISHING HOUSE,
Bairaut (Meerut).

Price Rs. 12-50 only

ALL RIGHTS RESERVED WITH THE AUTHORS

Printed at
SHRIHAL PRASS,
Meerut

PREFACE TO THE SECOND EDITION

The authors are thankful to the readers for the kind reception which they gave to the second edition. The book has been revised and a few more examples taken from recent Indian University Papers have been added at proper places. We hope, the present edition, is free from mistakes and misprints especially for fully revised chapters so that the text with this change will be found to be more useful to the readers.

We shall gratefully acknowledge the suggestions for the improvement of the book.

Baraut

Authors.

PREFACE TO THE FIRST EDITION

The text is intended mainly for the use of students offering one of the agriculture subjects (like Agronomy, Agr. Botany and Agr. Extension) at post graduate level in which Statistics is set as a compulsory subject. Many of these Students have a limited knowledge of mathematics. The present book, we believe, has a large number of exercises of all types firstly to satisfy the need of those for whom it is intended secondly to illustrate the theory amply.

The reader is expected to be familiar with the Arithmetic of school level and a limited knowledge of Algebraic symbols. The writers have purposely avoided proving the formulae lest it should unnecessarily burden the minds of the students.

It is particularly important that the reader should not form the impression that Statistical Methods contain a series of incomprehensible formulae to be applied indiscriminately to any available data. However, the care should be taken to grasp and appreciate the ideas and principles underlying the method learnt.

The book has been divided into three parts, namely,

(i) *Statistical Methods*,

(ii) *The Experimental Designs*,

and (iii) *Official Agricultural Statistics*,

and thus contains all the information and topics needed for M. Sc. Ag. Students. These parts have been further sub-divided into chapters. A large number of unsolved examples have been added at the end of each chapter for practice. Agra University Examination papers are also attached in the end of the book for M. Sc. Ag. Students of Agra University.

Writers of a text book are always indebted to all the earlier books on the subject and therefore we express our gratitude to all the publishers and writers whose books we have often consulted.

The authors are indebted to the staff members Prof. Fauran Singh, Prof. H. P. Singh and Prof. O. S. Verma, J. V. College, Baraut, for their critical comments and valuable suggestions which have undoubtedly improved the book. In particular, we consider it our noble duty to express our deep gratitude to reverend Gurujee prof. Mahendra Pratap, head of the Statistics Department, J. V. College, Baraut who has given his valuable guidance and assistance at all stages.

We are thankful to Prof. Rajendra Singh, head of the Statistics Department, A. S. Jat College, Lakhaoti (Bulan Shahar) for his assistance in calculations of the problems.

In spite of our best efforts, a number of misprints and mistakes are likely to have crept in a book of this type. We shall be grateful to any notice of these errors and for any suggestion for improving the book.

Baraut

Singh



Verma

CONTENTS

Part I

(Statistical Methods)

<i>Chapter</i>	<i>Subject</i>	<i>Page</i>
I.	Classification and Tabulation	1—25
II.	Graphs and Diagram	27—88
III.	Measures of Central Tendency and Dispersion	89—148
IV.	Elementary Idea of Probability	1—10
V.	Tests of Significance	11—57
VI.	Analysis of Variance	58—68
VII.	Correlation and Regression	69—114
VIII.	Sampling Techniques	115—128

Part II

(The Experimental Designs)

I.	Design of Experiments	1—14
II.	Completely Randomized Design	15—27
III.	Randomized Block Design	29—45
IV.	Latin Square Design	47—64
V.	Analysis of Covariance	65—92
VI.	Analysis of Incomplete Observations	93—117
VII.	Factorial Experiments.	119—151
VIII.	Confounding	153—166
IX.	Split Plot Design	167—180
X.	Switch Over Trials	181—186
XI.	Progeny Row Trials and Compact Family Block Design	187—192
XII.	Rotational Experiments	193—195

Part III

(Official Agricultural Statistics)

I.	Official Agricultural Statistics	1—3
II.	Land Utilization Statistics	4—8
III.	Method of Estimating Crop-yield	9—17

Appendix (1). t , χ^2 , F tables.

Appendix (2). University Examination Papers.

Part

Statistical Methods

Chapter 1

Classification & Tabulation

The raw statistical data which is collected in the course of an inquiry, usually consists of a series of readings and measurements. The original form of the collected data is very complex & voluminous and so it is very difficult to draw any inference from it as such. Thus, it is necessary to condense the data and present it in a systematic way. It is done by the 'Classification & Tabulation' of the data. By these processes, the volume and complexity of the data are reduced; which make the analysis and interpretations of the data easier.

Classification:—It is the process of arranging the individuals or items into groups or classes according to their similarities with respect to some variable character.

The Advantages of classification:—

(1) The whole data is divided into a number of classes as the items having resemblances with respect to certain character are put together in one class. Thus, it indicates the point of similarity and dis-similarity.

(2) The classification reduces the volume of the data which helps in forming the mental picture of the data.

(3) It brings into light those important informations which are liable to be ignored without classifying the data.

(4) It prepares the ground for comparisons and inferences. Because the classification of the data is not only sufficient for the comparisons and interpretations but it also helps in the tabulation of the data. Thereafter, the comparisons and the

interpretations on the data are possible.

(5) The classification provides the orderly arrangement of the items of the data.

Basis of classification:

The classification of the data is based upon the characteristic possessed by the items contained in the data. There are two types of characteristics:—

(i) Descriptive (ii) Numerical.

(i) **Descriptive-characteristic:** Under this type of characteristic, the quantitative measurements are not possible but only the presence or absence of a particular characteristic is observed in the terms available in the data. For example, sex, literacy, unemployment and blindness etc are the *Descriptive or qualitative characteristic* of the items.

Classification according to Attribute: The classification of the given data according to a descriptive characteristic is known as the classification according to attribute. Here only the presence or absence of a quality (attribute) forms the basis of classification

The classification can further be divided into two kinds;—

- (a) Simple classification or classification by dichotomy,
- (b) Manifold classification.

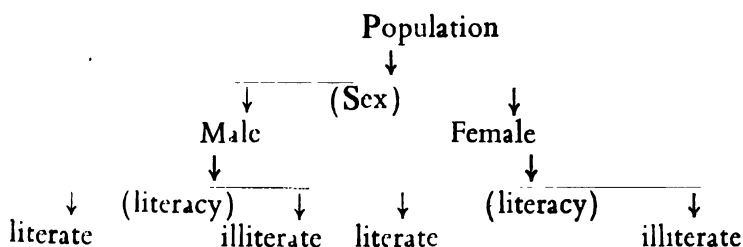
(a) **Simple Classification:** The kind of classification is termed as Simple classification where only one attribute is taken as the criterion of classification. Here, the whole data is divided into two classes:—

(1) Having those items which possess the attribute quality),

(2) Having those items which do not possess the attribute.

For example; the population of a country can be divided into two classes according to the attribute 'Unemployment' i.e. unemployed and employed or according to the attribute 'sex' i.e. male and female etc.

(b) **Manifold Classification** . The kind of classification is termed as Manifold classification where more than one attribute is taken as the criteria of classification. Here the whole data is divided into more than two classes. For example, the population of a country can be divided into four classes according to two attributes 'sex' & 'literacy' as shown in the diagram given below.



Thus the sub-divisions of the population are--

- (1) literate male (2) Illiterate male (3) literate female
(4) Illiterate female

Note:---

In classification according to attribute, it is necessary that the classes should be clearly defined before starting the actual work of classification i. e. a boundary should be set between the classes. As for example, in case of literacy we must decide in advance to whom we shall call the literate and to whom the illiterate before starting the work of classification of the data.

(ii) **Numerical Characteristic** : Under this type of characteristic, the quantitative measurements are possible. As prices, income, age, yield, height, and weight etc. are the examples of numerical characteristic.

Classification according to Class-Interval . The classification of the given data according to a numerical characteristic is known as the classification according to the class-interval.

In the classification according to class-interval, we must be ascertained before hand whether the data is discontinuous or a continuous one. The discrete variables are easier to deal with. An example of a discrete variable is the throws of a die.

Example(1): Let a die be thrown 50-times, and noted the number at its upper most face each time. If the number obtained are 1, 6, 5, 5, 4, 3, 1, 2, 3, 4, 2, 5, 6, 1, 2, 2, 2, 5, 1, 6, 6, 4, 6, 4, 3, 1, 6, 3, 1, 4, 3, 5, 6, 1, 3, 6, 4, 1, 1, 4, 3, 5, 3, 1, 2, 4, 4, 4, 2, 5, The variable (X) is the outcome of the number at the upper most face of the die in each throw and it clearly assumes the values 1, 2, 3, 4, 5, & 6.

In order to classify the data, these six values assumed by the variable (X) will be written in a horizontal row or a vertical column under the heading "*Variable Value (X)*". Since each value of X is repeated a number of times; the number of repetition of a variable value is called its frequency (f). In classification, these respective frequencies for each value of the variable are written against the values of the variable in the 2nd. column under the heading "*Frequency f* ". As in the present example, the variable value '1' is repeated 10 times and variable value '2' is repeated 7 times; so the respective frequencies for $X=1$ & 2 are 10 & 7.

Frequency Table : The table, showing the different values assumed by the variable and their respective frequencies is called a frequency table.

The frequency table for this given example is shown below:—

Table No. 1

Variable value (X)	Frequency (f)
1	10
2	7
3	8
4	10
5	7
6	8
Total	50

This method is adopted only, when the values taken by the variate are not so many (numerous).

On the other hand, when the number of values taken by the variate is large and the range between the greatest and the smallest value is also large, then the entire range is partitioned into classes of appropriate size by drawing the arbitrary lines showing the corresponding frequencies of the variate in each class. Each class is defined by two boundaries, the lower boundary and the upper boundary. The lower boundary of the class is called the *Lower Limit* of the class and the upper boundary is called the *Upper Limit* of the class.

Class-Interval: It is the difference between the upper & the lower limit of the class. For example, for the class 25—35, the upper limit is 35 and lower limit is 25 and the C. I. = $35 - 25 = 10$.

Variate Value : The mid-value of the class is called the

variate value i.e. $M.V. = \frac{\text{upper Limit} + \text{lower limit}}{2}$ and so in

the above mentioned class the $M.V. = \frac{35+25}{2} = 30$

Important points for frequency distribution :

While making a frequency distribution, the following points should be borne in mind—

(1) The class-intervals should be equal in each class to provide with easy computations.

(2) The number of classes should not ordinarily exceed 20 and it should not be less than 6. Because for more than 20 classes, the calculations become heavy and there is **not** much gain in the accuracy. Also, for less than 6 classes, there may be a great deal of loss in accuracy.

(3) In general, there should be no class without definite limits as under 'a' or over 'b'.

(4) We must treat all the values assigned to the different classes equal to the mid-value of the classes. Hence the class-interval must be chosen in such a way that the average of all the items in that class should not deviate much from the mid value of the class.

(5) A decision should be made before hand whether the class 'a-b' contains a lower limit of the class or b, upper limit of the class. A note of this should also be given, below the table.

(6) The class-interval should be an integer as far as possible.

Methods of Classification according to Class-interval:-

There are two methods of classification according to the class-interval.

(1) Exclusive Method (2) Inclusive Method

(1) **Exclusive Method**:—In this type of classification, the upper limit of any class is the lower limit of the succeeding class i.e. the class-limits are of the type a-b, b-c, c-d, etc. Hence, there is a confusion about those items which are exactly equal to the class-limits (called border line items). Usually, the border line items are placed

in the classes where they are as the lower limits of the class. As variable value 'b' is placed in the class b-c. This confusion can be removed by having the classes of the type, a and under b, b and under c and so on.

(2) Inclusive Method:—In this type of classification there is a gap between the upper limit of the class and the lower limit of the succeeding class. Hence, it leaves no room for confusion for the border line items. Thus the classification will be of the form a-b, c-d, e-f etc. and the class a-b will involve all the items from a to b inclusive a & b within it. This method can not be used in the case of a continuous variate.

Rules for making a frequency distribution:

(1) Before starting the actual work of classification, a preliminary inspection of the data should be made and the difference of the highest & the lowest values should be divided by the desired number of the class to be formed. It will give an approximate idea of the class-interval and will also help in setting the class limit.

(2) After deciding about the class-interval and class-limits, a table containing three headings namely, class-interval, tally marks & frequency should be prepared.

(3) Read for the items in the row table and for each items put the mark '1' in its corresponding class in the column headed *Tally marks*. After putting 4 marks of the above type the fifth is obtained by crossing the 4 marks and thus forming a group of 5 which helps in counting in the end.

(4) The sum of the tally marks is written in the column of frequency against the respective classes.

(5) The check of the sum of all the frequencies (Σf) must show Σf equal to the total number of variate-values.

Example (2)

The marks obtained by 60 students in the paper of statistics are given below—

22	47	9	42	31	17	13	15	18	25	35	33
38	15	0	33	10	34	29	26	16	33	16	27
24	22	26	19	14	36	18	25	21	12	21	10
28	25	17	38	10	3	31	24	3	2	36	18
26	29	27	39	28	35	26	27	13	33	25	18

Arrange the data in the form of a frequency distribution?

Sol:—

Here the variate is the numbers obtained in statistics and the original form of the data is an *ungrouped data*. If we arrange the data in the form of a table shown in the example (1), then there are as many as 29 values of the variable. Some of them occurring once, some twice, some thrice and some 4 times at the most.

Though, it is a frequency distribution giving a clear idea when compared to an ungrouped data, yet it may be improved by classifying it into groups. From an eye inspection of the data it is obvious that the range is $47-0=47$, because the highest value is 47 and the lowest value is 0. Further, the variable does not assume a fractional value and is a discrete variate having gaps between its successive values. The lowest value 0 suggests that we must start with lower limit as Zero and continue by 4 marks interval, adopting the Inclusive-Method of classification.

The frequency-table will be formed as given below—

Table No. 2

Class-interval	Tally marks	Frequency (f)
0—4		4
5—9		1
10—14	≡	7
15—19	≡ ≡	11
20—24	≡	6
25—29	≡ ≡ ≡	16
30—34	≡	7
35—39	≡	6
40—44		1
45—49		1
Totals		$\Sigma f = 60$

In a similar manner, we can form the frequency table (distribution) when the ungrouped data is such that it pertains to a continuous variate and there *Exclusive Method* of classification will be used.

Example (3)

The following table gives the Vickers Hardness number of 20 shell cases—

66.3	61.3	62.7	60.4	60.2
64.5	66.5	62.9	61.5	67.8
65.0	62.7	62.2	64.8	65.8
62.2	67.5	67.5	60.9	63.8

Arrange the data in the form of a frequency distribution?

Solution:—

The largest value is 67·8 and the lowest is 60·2, so the range is $67·8 - 60·2 = 7·6$; it suggests us to start with 60·0 as our lower limit and to continue by 1 (unit) class interval. The variable is the Hardness Number, and clearly it is a continuous variate.

The frequency-table is given below—

Table No 3

Class-interval	Tally marks	Frequency (f)
60—61		3
61—62		2
62—63		5
63—64		1
64—65		2
65—66		2
66—67		2
67—68		3
Totals	—	$20 = \Sigma f$

Cumulative Frequency Table : In some of the statistical investigations (Educational tests, wages or salary statistics etc.), we require the number of variates which are '*less than*' or '*more than*' a given value. For this purpose, it is necessary to change an ordinary frequency table into a cumulative frequency table. It can be done in the following two ways:—

(1) Less than type:

Suppose, we are given the daily expenses for one week as follows:—

Table No. 4

Days	Expenses (Rs.)
Monday	4
Tuesday	6
Wednesday	10
Thursday	16
Friday	12
Saturday	8
Sunday	4

Now if we want to know the total expenses upto Wednesday then it will be given by the sum of the expenses for Monday, Tuesday & Wednesday i.e. $4+6+10=20$. Similarly the total expenses incurred upto any day of the week can be obtained by adding the expenses from Monday upto that day. In the tabular form, the total expenses can be represented as given below in table No. 5.

Table No. 5

Days	Expenses up to the day (Rs.)
Monday	4
Tuesday	10
Wednesday	20
Thursday	36
Friday	48
Saturday	56
Sunday	60

On the other hand, the expenses after any day (including the day) will be obtained by adding the expenses for that day and after-wards. The following table No. 6 serves this purpose.

Table No. 6

Days	Expenses after the day (Rs.)
Monday	60
Tuesday	56
Wednesday	50
Thursday	40
Friday	24
Saturday	12
Sunday	4

It the days are replaced by the variate x , by its values x_1, x_2, \dots, x_7 , then the values in the 2nd. column of table No. 5 show the number 'less than' the variate value and column 2nd. of the table No. 6 shows the number 'more than' the variate value.

These frequencies, which are less than or more than the variate value (as shown in second column of table No. 5 & 6) are called the '*Cumulative Frequencies*'.

The similar procedure is adopted when instead of the variate value, the classes are given.

Example No. 4

Form a cumulative frequency table from the following data: —

class: 0-3 3-6 6-10 10-12 12-15 15-19 19-20 20-23 23-25
Frequency: 2 4 5 7 8 11 12 14 10

Solution: —

(i) In making the cumulative frequency table of "*Less than type*", the first class will be replaced by under 3, second class by under 6 and so on. The cumulative frequencies will be the sums of the frequencies upto that class for all the respective classes. The cumulative frequency tabale of "*Less than type*" is given below:—

Table No. 7

Variate value	Cumulative frequency (c. f.)
Under 3	2
Under 6	6
Under 10	11
Under 12	18
Under 15	26
Under 19	37
Under 20	49
Under 23	63
Under 25	73

(2) More than type:

(ii) In forming the cumulative frequency table of "*more than type*", the first class is replaced by more than 0, second by more than 3 and so on. The cumulative frequencies are the sums of frequencies onwards this class for all the respective classes. The cumulative frequency table of "*more than type*" is given below:—

Table No. 8

Variate Value	Cumulative frequency (c. f.)
More than 0	73
More than 3	71
More than 6	67
More than 10	62
More than 12	55
More than 15	47
More than 19	36
More than 20	24
More than 23	10

Note: (1)—

From the above example, it is clear that in the "Less than type" c. freq. table, we have to replace the classes by 'under' or 'below' the 'upper limits of the classes' and in the "More than type" c. freq. table we have to replace the classes by 'More than or over' the lower limits of the

c lasses'.

Forming an ordinary freq. table from a c. freq. table:—

There are the cases when we want to change the cumulative frequency table into an ordinary frequency table. For example, if for every unit of consumption of a certain commodity, its total utility is given and from it we want to know the marginal utility for every unit of consumption then it will be a case of forming an ordinary freq. table from a cumulative one.

Example (5)

In the table given below, the total utility for a unit consumption of a certain commodity is given. Find the marginal utility for every unit of consumption ?

Unit : 1 2 3 4 5 6 7 8 9 10 11 12

Total utility : 3 11 21 33 49 63 73 81 98 103 107 108

Solution:—

We know that:—

the total utility for i th unit = the marginal utility for 1st unit
+ ... + the marginal utility for i^{th} unit.

From this, it is clear that the relation between the total utility and the marginal utility is the same as between the cumulative frequency and the ordinary frequencies.

Here we have to convert a cumulative frequency table into an ordinary frequency table as shown below:—

Table No. 9

Unit	Marginal utility	Unit	Marginal utility
1	3	7	10
2	8	8	8
3	10	9	17
4	12	10	5
5	16	11	4
6	14	12	1

Note: (2)—

To form the ordinary frequency table from cumulative frequency table of “less than type” we have—

frequency for i^{th} class = c. f. for i^{th} class — c. f. for $(i-1)^{\text{th}}$ class.

Similarly, to convert a c. f. table of “more than type” into an ordinary frequency table we have

frequency for i^{th} class = c. f. for $(i)^{\text{th}}$ class — c. f. for $(i+1)^{\text{th}}$ class.

Example (6)

From the following data, find the number of persons in the classes 20-25, 25-30..... and 55-60.

Age more than :	20	25	30	35	40	45	50	55
No. of persons :	800	750	680	580	400	250	130	60

Solution:—

We have been given that the number of persons who are old than 20 is 800 and those who are old than 25 is 750. Thus the no. of persons who are between the age of 20 & 25 is $800-750=50$ only. In the same way, the no. of persons between 25-30, 30-35 and 55-60 can be found. Because the formula used is:—

frequency for i^{th} class = c. f. for i^{th} class — c. f. for $(i+1)^{\text{th}}$ class.

Thus the table on the next page shows the no. of persons between the different age-groups.

Table No. 10

Age-group	No. of persons
20—25	50
25—30	70
30—35	100
35—40	180
40—45	150
45—50	120
50—55	70
55—60	60
Total	800

Example (7)

Obtain the ordinary frequency table from each of the following tables:—

(1)				(2)			
		No. of Students				No. of Students	
Marks Below	10	3		Marks above	0	30	
" "	20	8		" "	10	26	
" "	30	17		" "	20	21	
" "	40	20		" "	30	14	
" "	50	22		" "	40	10	
				" "	50	0	

Solution:—

(1) Here the c. f. table of 'Less than type' is given and we have to convert it into an ordinary frequency table. The frequency for i^{th} class = c. f. for i^{th} class - c. f. for $(i-1)^{\text{th}}$ class. Thus, the table given below shows the desired no. of students in the different groups of marks:—

Table No. 11

rks	No. of students
0—10	3
10—20	5
20—30	9
30—40	3
40—50	2
Total	22

(2) Here the c. f table of “More than type” is given and we have to convert it into an ordinary frequency table. The frequency for i^{th} class = c. f. for i^{th} class — c. f. for $(i+1)^{\text{th}}$ class. Thus, the table given below shows the desired no. of students in the different groups of marks:—

Table No. 12

Marks	No. of students
0—10	4
10—20	5
20—30	7
30—40	4
40—50	10
Total	30

Tabulation: —

It has been mentioned already that the classification alone is not sufficient for comparisons and inferences. The data is further put to the treatment of tabulation which make it fit for further analysis and comparisons.

Thus the *Tabulation* is the process of arranging the classified data in an orderly manner into rows & columns in such a way that it makes clear all the important informations contained in the data.

For tabulation, the following points should be kept in mind—

(1) Every table should have a title, explanatory in itself and it should provide the informations regarding:—

- (a) What the data are. (b) Where the data are.
- (c) Principle of classification. (d) Time of data.

(2) The rows & columns must be arranged in a logical order to facilitate the comparisons.

The headings & sub-heading should be concise and without any ambiguity.

(4) The units of the data presented, should be clear.

(5) The complicated tables should be avoided and so they must be broken up into parts, in more than one simple tables.

(6) The size of the table should suit the size of the paper.

(7) The table should be presented in a neat, clean and comprehensive way by drawing the double lines wherever required

(8) The source of the table should also be written.

Types of Tabulation

There are two types of tabulation:—

- (1) Simple tabulation (2) Complex tabulation.

Simple Tabulation:—

In a simple tabulation, the data is presented with respect to one character only or in other words, a simple table is capable to provide the information relating to one character only. For example, the population of India can be divided

according to religion. The table for this purpose will be as follows:—

Table No. 13

Religion	Population
1. Hindu	...
2. Muslim
3. Sikh	...
4. Christian	...
5. Jain	...
6. Others	...
Total	...

(2) Complex Tabulation:—

A complex table gives the information relating to several characters. This type of tabulation is further classified as 'Double Tabulation' 'Triple or Treble Tabulation' and 'Manifold Tabulation'.

Double or two fold Tabulation:

In this type of tabulation, the data is divided according to two characteristics. For example, the population of India can be divided according to religion & States and the table for it is given below:—

Table No. 14

States	Population						Total
	Hindu	Muslim	Sikh	Christian	Jain	others	
1. Assam							
2. Bihar							
3. Punjab							
4. U. P.							
.. .. .							
.....							
Total				..			

Treble Tabulation :

In this type of tabulation, the data is sub-divided according to three characteristics. Treble tabulation is capable of answering three mutually dependent questions. The table given below is an example of treble tabulation. It shows the distribution of India's population according to States, religion and sex.

Table No. 15

States	Population											
	Hindu		Muslim		Sikh		Others		Totals			
	Male	Female	Total	Male	Female	Total	Male	Female	Total	Male	Female	Total
1. Assam												
2. Bihar												
3. Punjab												
4. U P.												
.....												
.....												
.....												
.....												
Total												

Manifold Tabulation :

In this type of tabulation, the data is divided with respect to more than three characteristics. From this type of table, we can get the information relating to several (more than three) characteristics. An example of this type of tabulation is the division of India's population according to States, religion, sex, literacy and age etc.

Example (8)

Draw up in detail, with proper attention to spacing, double lines etc. and showing all the sub-totals, a blank table, in which could be entered the numbers occupied in the six industries at two dates (July 63, July 64) distinguishing males

from females and among the latter, single, married and widowed.

Solution:—

The desired table is given below:—

Table No. 16

Industry	No. in July, 63				No. in July, 64				Totals				Totals
	Female			Total	Female			Total	Female			Total	
	Male	Widowed	Married		Single	Male	Widowed		Married	Single			
1.													
2.													
3.													
4.													
5.													
6.													
Total													

The above table shows the no. of employees in six industries according to their distribution into sex and two dates.

Example (9)

Draw up two independent blank tables, giving rows, columns, and totals in each case, summarizing the details about the number of families, distinguishing males from females, earners from dependents and adults from the children.

Chapter II

Graphs And Diagrams

After condensing and summarizing the complex and numerous data in a systematic manner by means of classification and tabulation, we require certain devices which may present the condensed form of the data. This condensed form of the data is such that it becomes at once comparable and leaves an everlasting impression on the brain of the observer. One of the methods making the data intelligible is to represent it by means of graphs and diagrams. The graphic & diagrammatic representation of the data is always appealing to the eye as well as to the mind of the observer.

Advantages of Graphs & Diagrams:—

(1) The graphs & diagrams represent the data in attractive and appealing way both to the eye and mind.

(2) The graphic & diagrammatic representation of the data leaves an everlasting effect on the brain.

(3) The diagrams are not only attractive and impressive but they save time also. Because through the diagrams, it is possible to have an immediate grasp of significance.

(4) Another merit of the diagrams is the ease with which the two sets of the data are compared with each other.

(5) Forecasting becomes easier with the help of the graphs.

(6) The graphs are helpful in the interpolation also and they give an indication of correlation between the two variables.

(7) The partitioning values (median & quartiles etc.)

and mode are also determined by graphs.

Limitations of graphs & diagrams:—

(1) The graphs and diagrams provide an approximate picture of the data and not the accurate one. Thus they are useful in representation to general public but not to a statistician or an investigator who is interested in the detailed study.

(2) They are used to compare only such data which are technically comparable.

(3) They are easily capable of misuse.

Graphical Representation of Frequency Distributions:

In the graphical representation of the frequency distributions the horizontal axis of the graph is used to show the variate values and the vertical axis for the frequencies of the variate. The graphs of the frequency distributions are of the following types—

(1) Histogram (2) Frequency-polygon and Frequency Curve (3) Cumulative frequency curve or ogive.

(1) **Histogram** : It is composed of a set of rectangles one over each class-interval on the X-axis. The width of the rectangle is taken proportional to the class-interval and its height is taken proportional to the frequencies in the case of equal intervals. But in the case of unequal class-intervals the height of a rectangle is taken proportional to the ratio of frequencies to the class-interval and thus the area of the rectangle is proportional to the frequencies of the variate or the classes.

Example (1) :—

The following is the frequency distribution of the yield of sugar-cane in tons per acre.

Class :	35-40	40-45	45-50	50-55	55-60	60-65	65-70
Frequency :	7	8	12	26	32	15	9

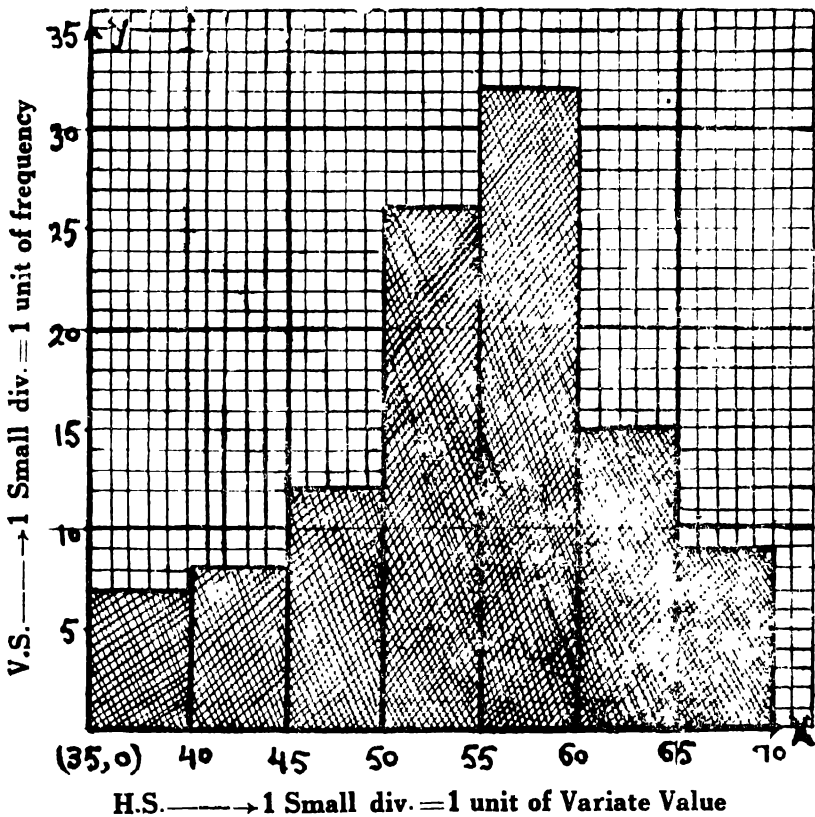
Drawn a histogram representing the above distribution ?

Solution:—

In this example, the class intervals are equal and so the heights of the rectangles will be proportional to the frequencies.

Graph No. 1 **HISTOGRAM**

“representing the yield of sugarcane in tons/acre”



Example No. (2): —

Draw the histogram of the following distribution:—

Heights : 48-50 50-54 54-56 56-58 58-59 59-60 60-63
(in inches)

No. of Boys :

(frequency) : 12 60 20 16 5 2 3

Solution :—

In this example, the class-intervals are not equal, so the heights of the rectangles will be proportional to the ratio of the frequencies to their respective class-intervals. Thus the

heights of the rectangles are given in the following table against the respective classes:—

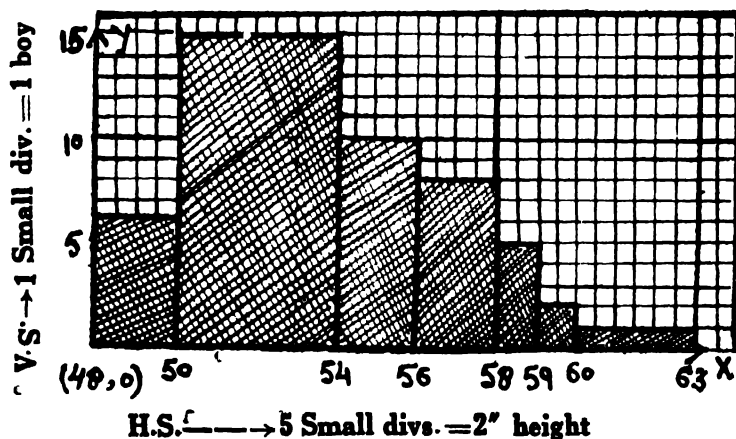
Table No. 18

Classes	Frequency/Class Interval
48—50	$12/2 = 6$
50—54	$60/4 = 15$
54—56	$20/2 = 10$
56—58	$16/2 = 8$
58—59	$5/1 = 5$
59—60	$2/1 = 2$
60—63	$3/3 = 1$

Graphs No. 2

HISTOGRAM

“representing the heights of the boys”



Frequency Polygon & Frequency Curve:—The frequency polygon is constructed by joining the points by means of straight lines whose abscissae are the mid points of the classes and the ordinates are the corresponding frequencies. A frequency polygon can also be obtained by joining the mid-points of the upper sides of a histogram by straight lines.

A frequency curve is a smooth, free hand curve, drawn through all the points which are obtained for the frequency polygon. The area under the curve should be equal to that of the histogram. Frequency polygon is used to find the 'mode' which is the *apex* of the curve.

Example (3a) —

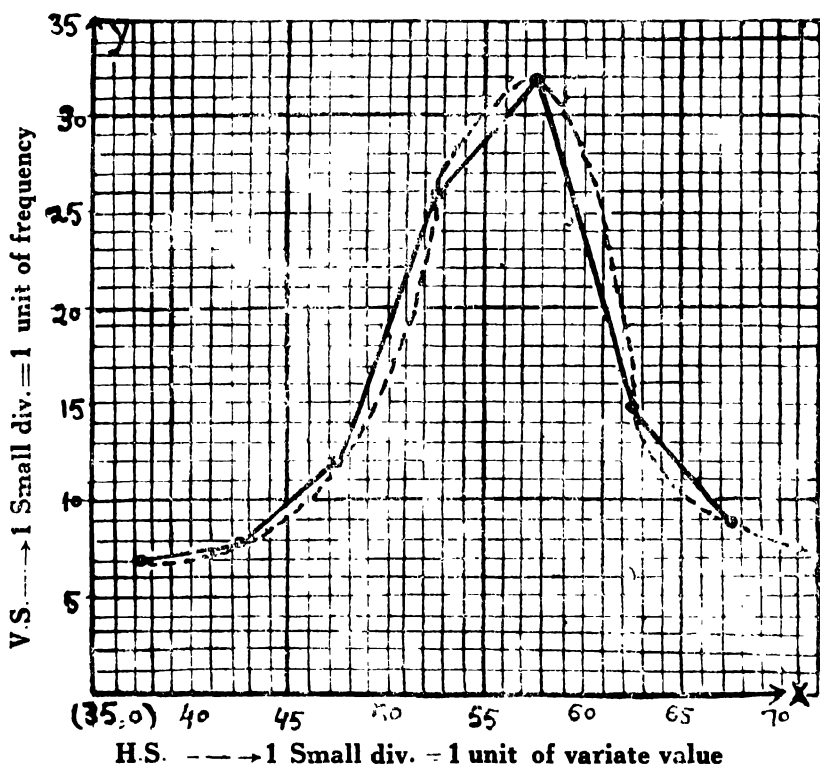
Draw a frequency ploygon & a frequency curve for the data in example (1) on page 28.

Solution: —

In this example, the points of the classes are plotted on the X-axis and the frequencies on the Y-axis.

Graph No. 3 (a)

Representing the yield of sugar cane in tons/acre.



Example 3 (b):—

The following table gives the distribution of height in inches for 100 students.

Interval	Frequency
>57 upto 60	3
>60 „ 63	12
>63 „ 66	31
>66 „ 69	37
>69 „ 72	16
>72 „ 75	1
Total	100

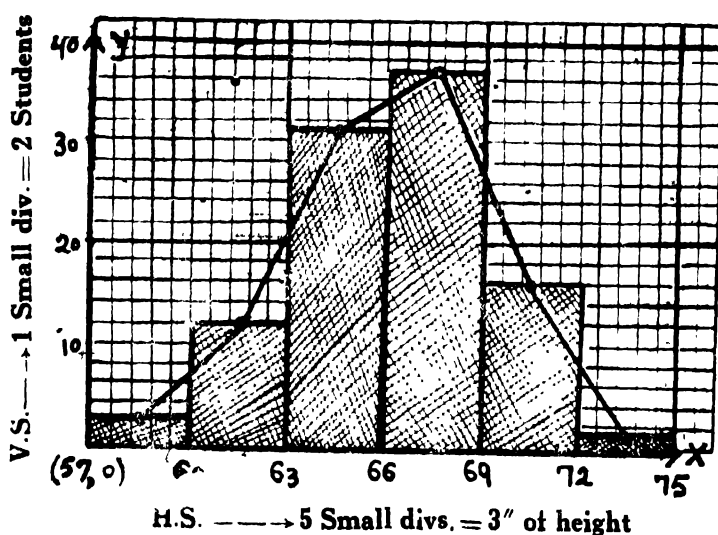
Represent the data in the form of a histogram as well as a frequency polygon ?

Solution:—

See the graph No. 3 (b)

Graph No. 3 (b)**HISTOGRAM & F. POLYGON.**

Representing the distribut on of height for 100 students



(3) **Cumulative frequency curve or ogive** : The ogive is constructed from a less-than type cumulative frequency table. The upper limits of the classes are taken as abscissae and the corresponding cumulative frequencies as ordinates and thus all the points are plotted on the graph. The free hand smooth curve through all these points is called the *ogive*. Ogive is helpful in determining the partitioning values (median, quartiles etc.).

Example (4):—

Draw a cumulative frequency curve for the data given in example (1) on page 28.

Solution:—

First of all, we shall construct the cumulative frequency table from the given data as follow:—

Table No. 19

Variate Value Under	Cumulative frequency(c.f.)
Under 40	7
Under 45	15
Under 50	27
Under 55	53
Under 60	85
Under 65	100
Under 70	109

The cumulative frequency curve is shown below:—

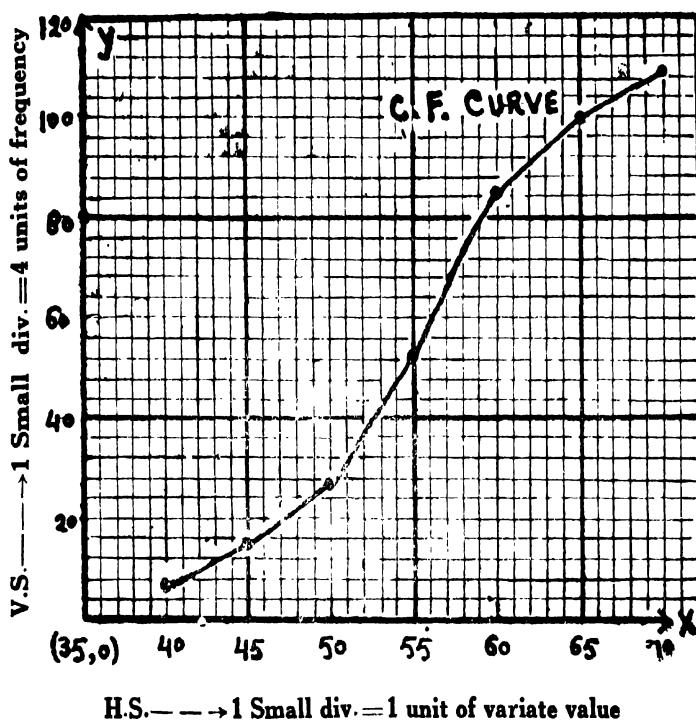
32

Exi

inc

Graphs No. 4

Representing the yield of sugar-cane in tnos/acre



Gap between the upper & the lower limits of the adjacent classes:—

In the case of the grouped data of the type a-b, b-c, c-d, and so on, we have seen that the graphs of the above three types can easily be used. But in case, where there is a

gap between the upper and the lower limits of the adjacent classes, the class limits should be so modified that the gap vanishes and the upper & lower limits of the adjacent classes become the same.

Example (5):—

Draw the histogram, frequency Polygon and the ogive for the following data:—

Scores :	20-29	30-39	40-49	50-59	60-69	70-79
No of students :	2	14	22	20	14	3

Solution:—

In this example, there is a gap between the upper & the lower limits of the adjacent classes. Hence, first we modify the classes to have the continuous form of the data and thus finish the gap. For the construction of histogram, frequency polygon & the cumulative frequency curve for this data, we form the following table:—

Table No. 20

Classes	Frequency (f)	Cumulative frequency (c.f)
19.5—29.5	2	2
29.5—39.5	14	16
39.5—49.5	22	38
49.5—59.5	20	58
59.5—69.5	14	72
69.5—79.5	3	75

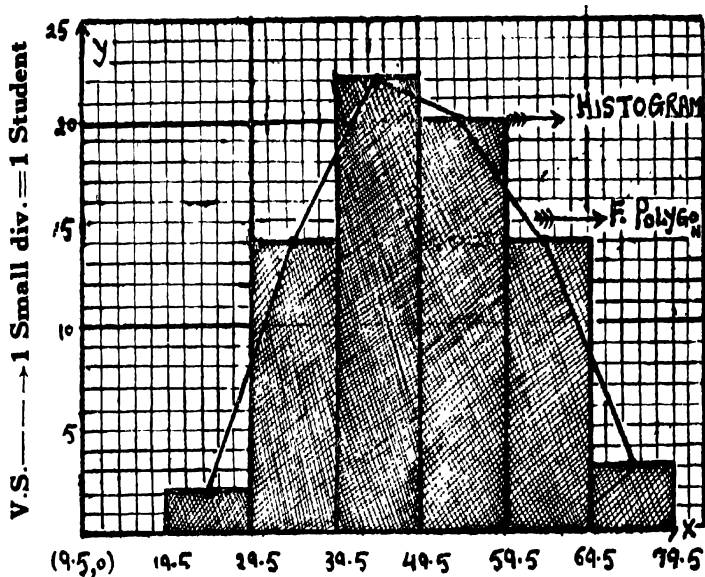
Graph No. 5 (a)**HISTOGRAM & F. POLYGON**

"representing the scores for Students"

32

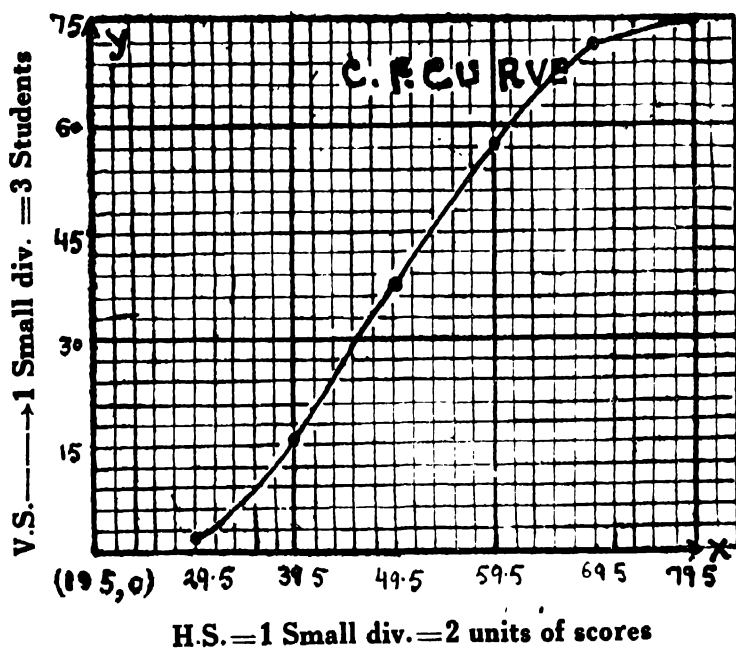
Ext

inc



H.S. ———→ 1 Small div. = 2 units of scores

Graph No: 5 (b)
Cumulative-Frequency-Curve
Or
OGIVE



32

E₃

ir

Graph For Discrete Variable:—When the data is given for a discontinuous variable accompanied by its corresponding frequencies, then the graphical representation of such a data is made by drawing thick lines parallel to the Y axis. These lines are drawn at the points of the discrete values of the variable shown on X-axis and their heights are proportional to their corresponding frequencies.

Example No. (6)

Represent the following frequency-distribution graphically:—

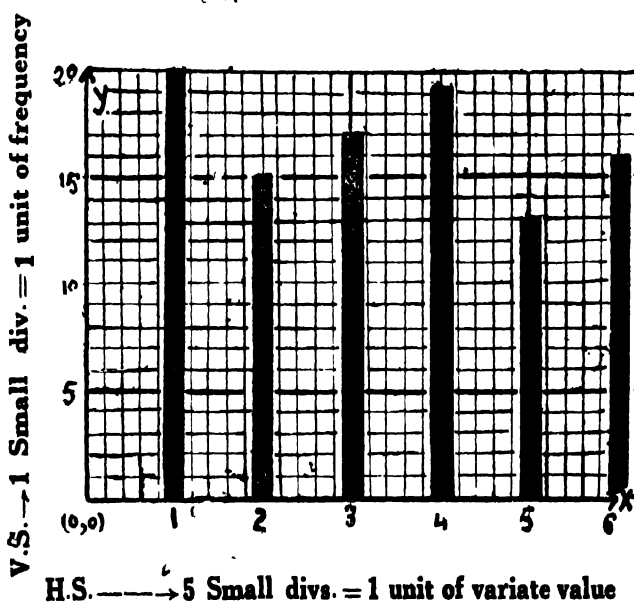
X :	1	2	3	4	5	6
f :	20	15	17	19	13	16

Solution:--

In the graph, the variate values are shown on the X axis and their corresponding frequencies on the Y-axis.

Graph No. 6

“Graph for discrete variable”



Graph for Time Series : (Historigram) The Historigram is the graph obtained from the time series data. In these graphs, the time variable 't' is measured along X-axis and its corresponding values ' u_t ' along Y-axis. All the available points form the data are plotted on the graph and joined by means of straight lines.

Following few examples will illustrate the type of graph:—

Example No. (7): --

The following table provides the yearly figures of production of sugar in lakh tons—

Year	production of sugar (lakh tons)
t	u_t
1948	9.65
1949	8.90
1950	6.75
1951	10.50
1952	11.25
1953	10.70
1954	12.35
1955	12.55
1956	12.95
1957	13.00
1958	13.50

(i) Draw the graph to show the fluctuations of the production of sugar.

(ii) Comment on the fluctuations of the production.

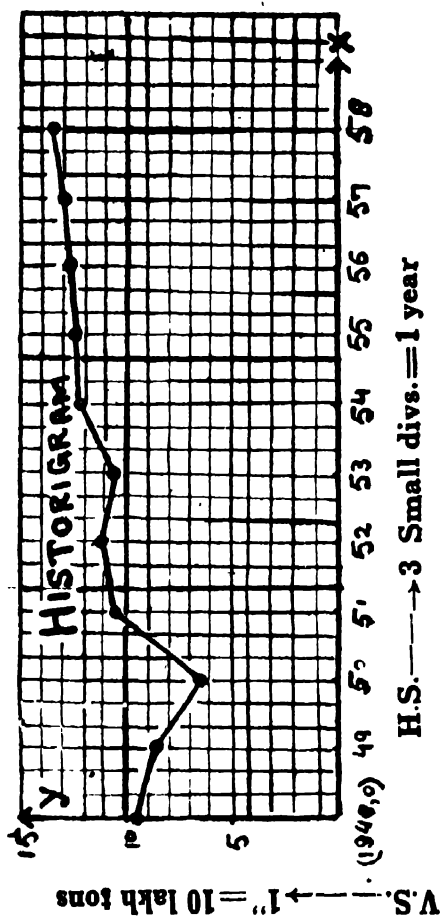
Solution:—

(i) The production of the sugar in lakh tons is taken along Y-axis and years on the X-axis in the following graph.

35

F

Graph No. (7)
Representing the production of sugar in lakh tons, from 1848—58



Example No (8):—

Represent the figures given below, on a graph-paper and comment on trend shown by the data.

Year	Price for Arhar in Rs./maund.
1929	4.0
1930	4.6
1931	3.6
1932	3.6
1933	3.3
1934	3.3
1935	4.7
1936	3.4
1937	4.3
1938	4.3
1939	4.2
1940	3.9

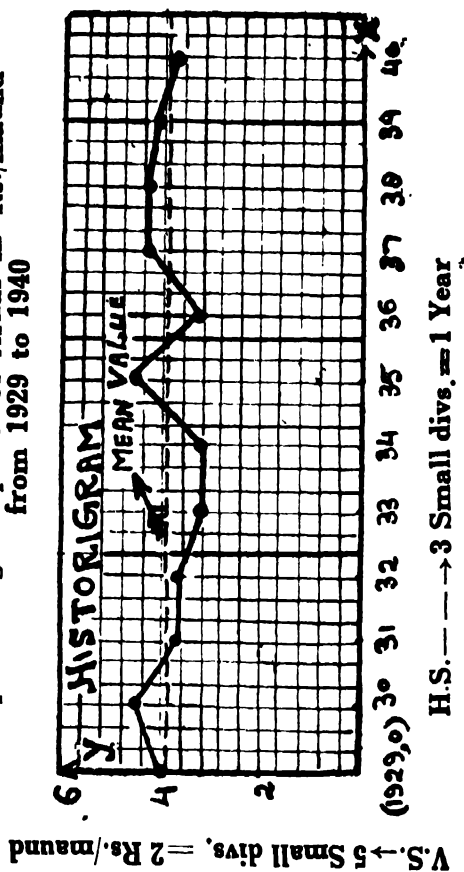
Solution :

An eye inspection of the graph shows that the prices are neither decreasing nor increasing but they are fluctuating around the price Rs. 3.9/maund, the mean value.

35

F

Graph No. (8)
Representing the price for Arhar in Rs./maund
from 1929 to 1940



Example No. (9):—

Draw the graph from the following table, in which the growth of the three kinds of wheat may be read from the graph. Discuss the growth of the three kinds of wheat ?

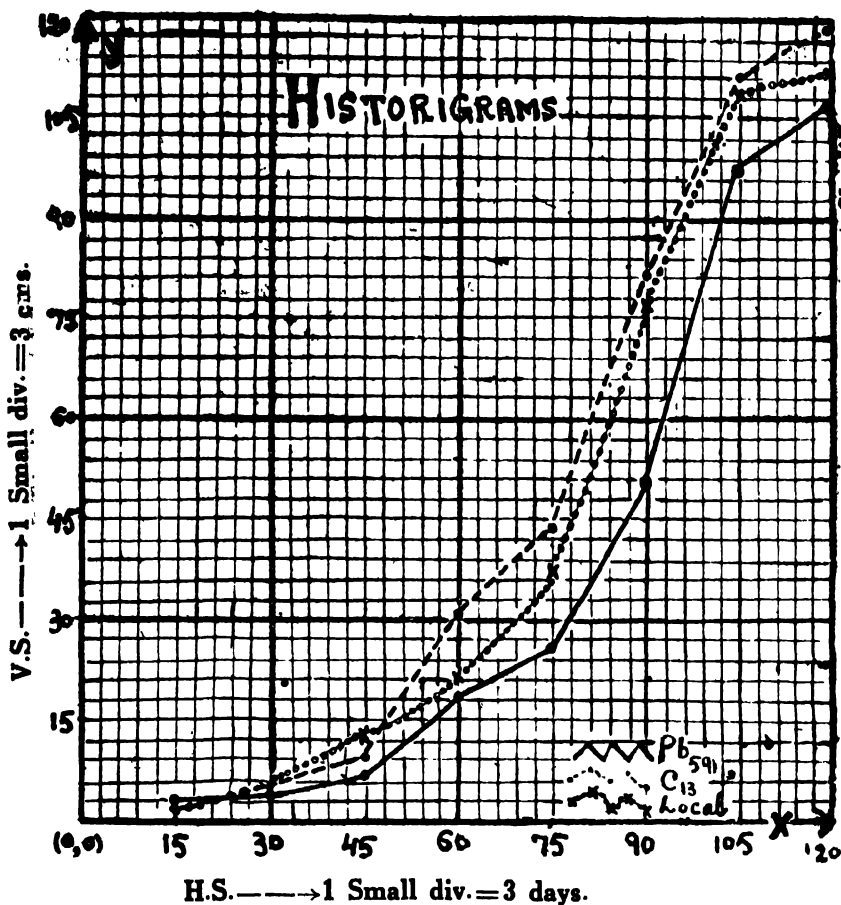
(U. P. Board 1956)

Age Days Kinds	Height in centimeters							
	15	30	45	60	75	90	105	120
Pb ₅₉₁	3.0	3.5	7.0	19.0	26.0	51.0	98.0	106.5
C ₁₃	2.5	4.0	10.0	31.0	44.0	82.0	111.5	119.5
Local	3.0	4.5	12.5	22.0	37.5	77.5	109.0	112.5

Solution:—

Of all the three varieties of wheat, the growth in C₁₃ is minimum upto the age of 45 days in comparison to the local variety and after this, the growth of C₁₃ is the highest, while that of Pb₅₉₁ is minimum. The growth of the local variety is between the two.

Graph (9)
Representing the growth of three kinds of wheat



An inspection of the graph clearly shows that upto the [age of 45 days, the growth in local variety is the highest and after it that of C_{13} is the highest. Thus, as regards the character of growth, C_{13} is the best.

Example No. (10):—

The average yields in maunds per acre of rice and wheat crops in a state of India since 1950-51 are given below. Comment whether the yields are increasing or decreasing ?

Year	Rice	Wheat
1950—51	5·27	8·88
1951—52	4·38	8·25
1952—53	5·31	9·21
1953—54	6·46	9·15
1954—55	6·12	9·64
1955—56	7·18	8·26
1956—57	6·21	8·46
1957—58	6·19	7·89
1958—59	7·13	8·61
1959—60	6·15	9·21
1960—61	7·15	10·91

(M.Sc. Ag. 1962)

Solution:—

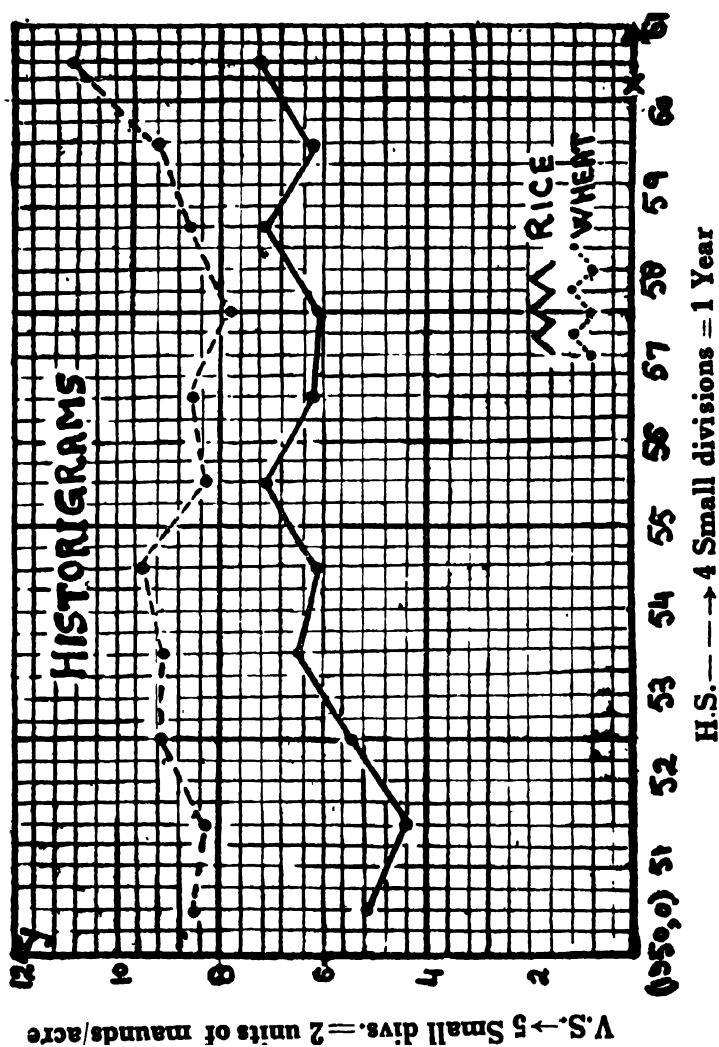
Let us first plot the yields of rice and wheat since 1950—51 to 1960—61 on the graph-paper. As a result of this, we obtain the following graphs.

An eye-inspection of the above graphs tell us that—
(i) the yield of wheat has an increasing tendency. The yield from 1950—51 to 1954—55 is increasing except a slight downward fluctuation for the year 1951—52. Further the yield is decreasing upto 1957—58 and again increasing for the remaining years.

(ii) The yield of rice is increasing from 1950—51 to 1955—56 except a slight downward fluctuation for the year 1954—55. For the next two years, the yields are decreasing and are the same. For the year 1958—59, the yield increases and again decreases for the year 1959—60 and increases for 1960—61. Thus on the whole, we can say that the yield shows an increasing trend.

Graph No. (10)

Representing the yield of Rice & wheat from (1950-51) to (60-61)



Example No. (11):—

The following table provides the yearly figures of production in lakh tons in U. P. Comment on the wheat production in U. P.

Year	Yield	Year	Yield
1943—44	19.01	1952—53	28.24
1944—45	23.56	1953—54	31.06
1945—46	22.83	1954—55	32.84
1946—47	23.37	1955—56	30.41
1947—48	21.94	1956—57	31.15
1948—49	19.74	1957—58	27.06
1949—50	24.26	1958—59	30.36
1950—51	26.78	1959—60	32.42
1951—52	25.33	1960—61	38.82

(M.Sc. Ag. 1963)

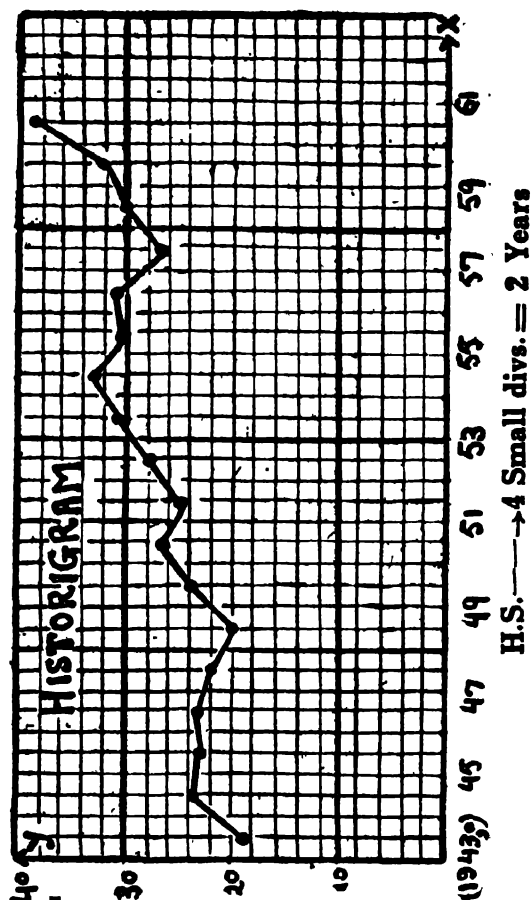
Solution:

The yields of wheat in U. P. since 1943—44 to 1960—61 are plotted on the graph-paper as shown below.

The graph shows that the yields are increasing upto 1945—46 and further decreasing for the next two years. Again from 1948—49 to 1954—55 the yields are increasing except a downward fluctuation for the year 1951—52. After this, the yield neither shows an increasing nor a decreasing tendency, it merely fluctuates around the yield 30 lakh tons. But for the year 1960—61, the yield increases. Thus on the whole, there is an increasing trend in the production of wheat.

Graph No. (11)
Representing the production of wheat in lakh tons from
(1943-44) to (1960-61)

V.S. \rightarrow 1 Small div. = 2 lakh tons



Diagrams

The following are the important diagrams—

- (1) One dimensional diagrams,
- (2) Two dimensional diagrams,
- (3) Three dimensional diagrams,
- (4) Cartograms and
- (5) Pictograms.

(1) **One Dimensional Diagrams:**—The one dimensional diagrams are the lines or bars arranged in ascending or descending order of magnitudes (length) along a vertical or a horizontal scale. The lengths of the bars are proportional to the magnitudes of the items.

(2) **Two Dimensional Diagrams:**—In the two dimensional diagrams, the magnitudes of the items are represented by the areas as by squares, rectangles and circles etc

(3) **Three Dimensional Diagrams:**—In the three dimensional diagrams, the data are represented by the cubes or cylinders and the magnitudes of the items are represented by the volumes of the cuboids or cylinders.

(4) **Cartograms:**—Here the data is presented by means of maps.

(5) **Pictograms:**—Here the data is presented by means of pictures.

But the bars and circular diagrams are most commonly used because of their accuracy and easiness in sketching.

“One Dimensional Diagrams”

The bar diagrams of common use in the one dimensional diagrams are of the following main types:—

- (1) Simple bar diagrams,
- (2) Multiple bar diagrams,
- (3) Sub-divided bar diagrams.

(1) **Simple bar diagrams:**—Here the magnitudes of the items are represented by thick bars of uniform width with equal spacing between any two consecutive bars. The lengths of the bars are proportional to the magnitudes of the items.

One bar is drawn for each item and they are arranged in the ascending or descending order of lengths along a vertical or horizontal line. These simple bar diagrams are appropriate for the data contained in simple tables.

Example (12)

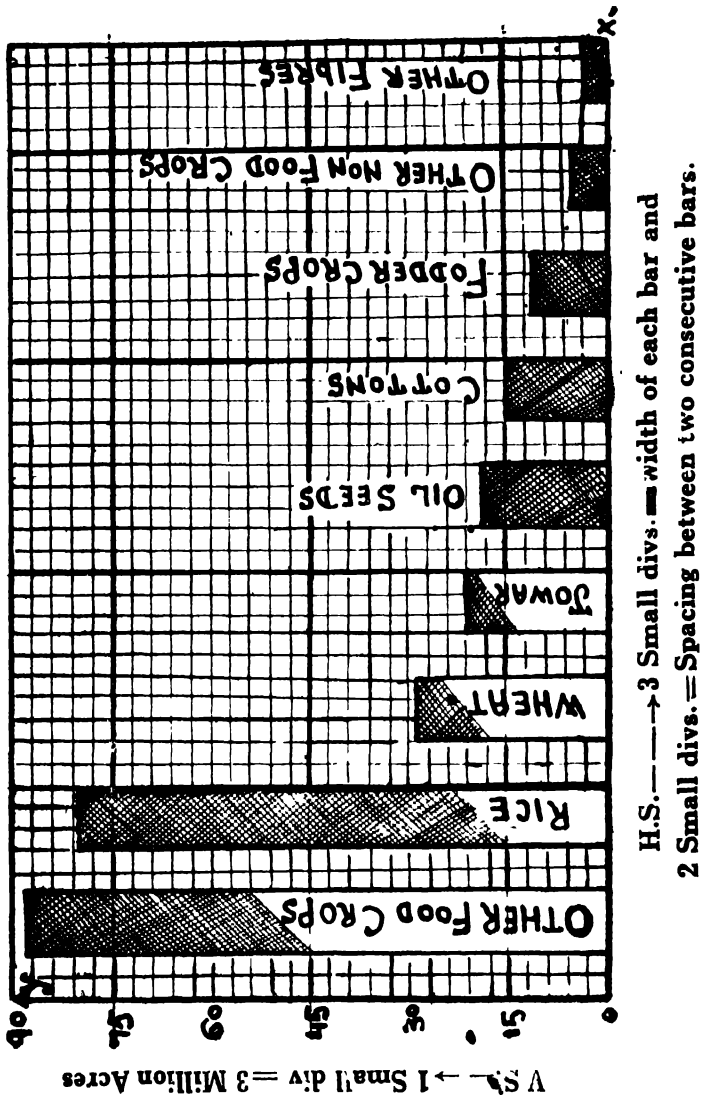
Draw a simple diagram to represent the following statistics relating to the area under different crops:—

Crops.	Million Acres
Rice	80·3
Wheat	27·6
Jowar	21·4
Other food crops	88·2
Oil seeds	17·6
Cotton	14·5
Other fibres	3·1
Fodder crops	10·2
Other non food crops	3·9

Solution:—

The Acreage (million acres) under different crops is shown by the simple bar diagrams:—

Diagram No. 1
“Simple Bar Diagrams for Acreage under different Corps”



(2) **Multiple Bar Diagrams:**—These bar diagrams are constructed like simple bar diagrams. They represent more than one type of data at a time and so two or more bars (as the case may be) are constructed at a time side by side.

Example 13 (a)

The following table gives the number of motor-cars produced in the three countries during the period 1929 – 1935.

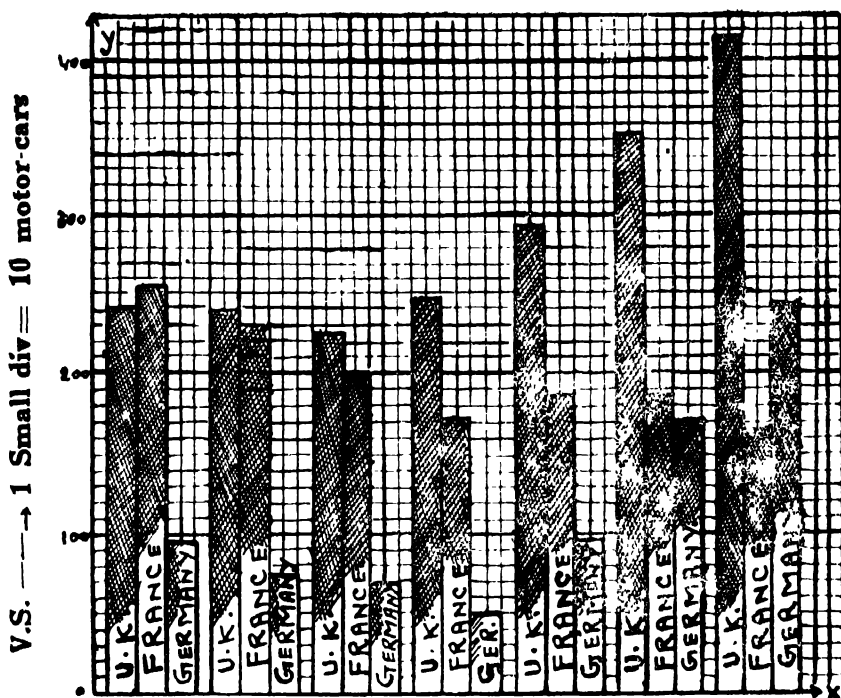
Year	Germany	France	U.K.
1929	96	254	241
1930	74	231	241
1931	68	201	226
1932	50	172	248
1933	99	189	296
1934	172	187	355
1935	245	166	417

Solution:—

Taking three bars at a time, side by side, we completed the diagram. Here the time (year) is taken along the X-axis and the number of cars produced in these countries along the Y-axis.

Diagram No. 2 (a)

"Multiple Bar Diagram for No. of Cars
Produced in U.K., France & Germany"



1929 1930 1931 1932 1933 1934 1935

H.S. ———→ 6 Small divs. = 1 year ,

2 Small divs. = width of each bar

1 Small div. = spacing between each set of three bars

Example No. 13 (b):—

The following table provides the percentage of cultivators and percentage of cultivated area in different sizes of holdings in U.P. Depict the data in a bar diagram to the scale ?

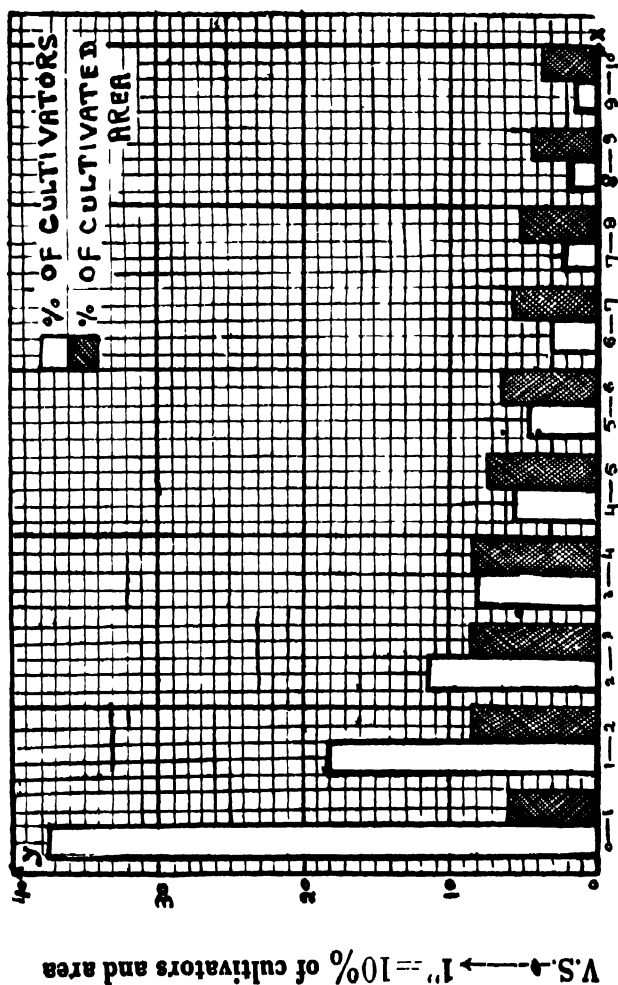
(M.Sc. Ag. 1964)

Size of holding (acres)	Percentage of cultivators	Percentage of area
Up to 1 ..	37.8	6.0
1—2	18.0	8.1
2—3.....	11.6	8.7
3—4.....	8.1	8.4
4—5	5.7	7.6
5—6	4.2	6.8
6—7	3.0	5.9
7—8.....	2.3	5.1
8—9	1.8	4.4
9—10.....	1.4	3.9

Solution:—

The percentage of cultivators and the cultivated area in different sizes of holdings will be depicted by the compound bar-diagram. Corresponding to each holding size, two adjacent bars will be constructed where one of the two will represent the % of cultivators and the other % of cultivated area.

Diagram No: 2 (b)



H.S. ———→ 4 Small divs — 1 acre (size of holding)

2 " " " " Width of each bar

1 " " " " Spacing between each pair of bars

V.S. ———→ 1" = 10% of cultivators and area

(3) **Sub-divided Bar Diagrams:**—These diagrams are suitable for the data given in the complex tables; where the magnitudes of the items are sub-divided into sub-classes.

Example (14)

The following table gives the populations of males out of the total populations for different years for a district. Draw a suitable diagram to show the populations given in the data.

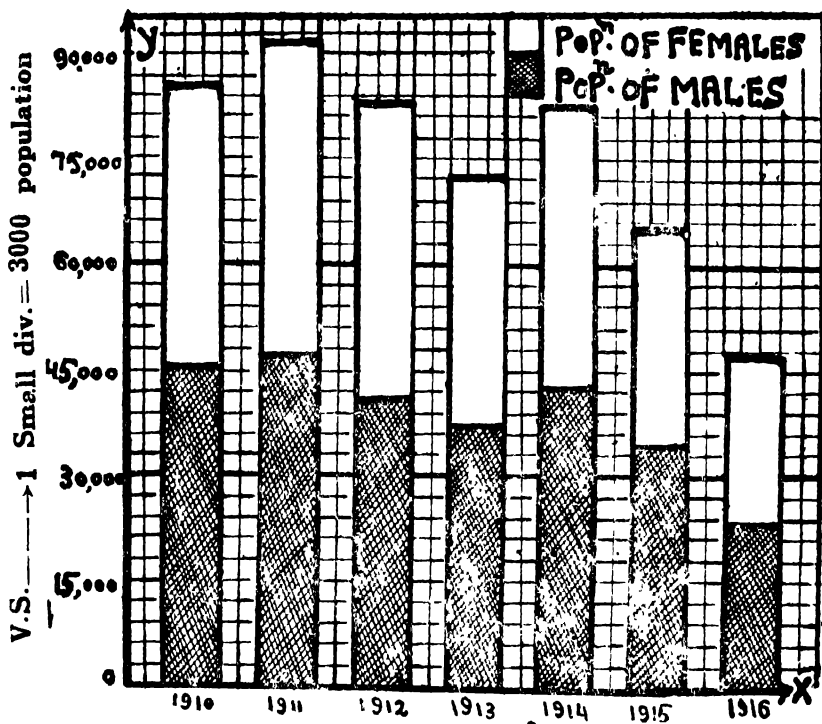
Year	Total Population	Population of Males
1910	85761	45761
1911	92821	47212
1912	83728	41312
1913	72511	37256
1914	83123	42218
1915	65735	34725
1916	46849	23428

Solution:—

The bars are constructed for different years and their lengths are proportional to the magnitudes of the total populations. Then each bar is sub-divided to show the populations of males for the different years.

Diagram No. 3

"Sub-divided bar diagram for total population
& population of males"
from (1910-16)



H.S. ———> 3 Small divs. = 1 year

= width of each bar

2 Small divs. = Spacing between two consecutive bars

Example (15):—

The following table shows the expenditure incurred by the central government and the governments of States of type A & B in the 1st five years-plan under the six major heads. Represent the data by means of a suitable diagram :

Expenditure in the 1st five years-plan (in crores of rupees)

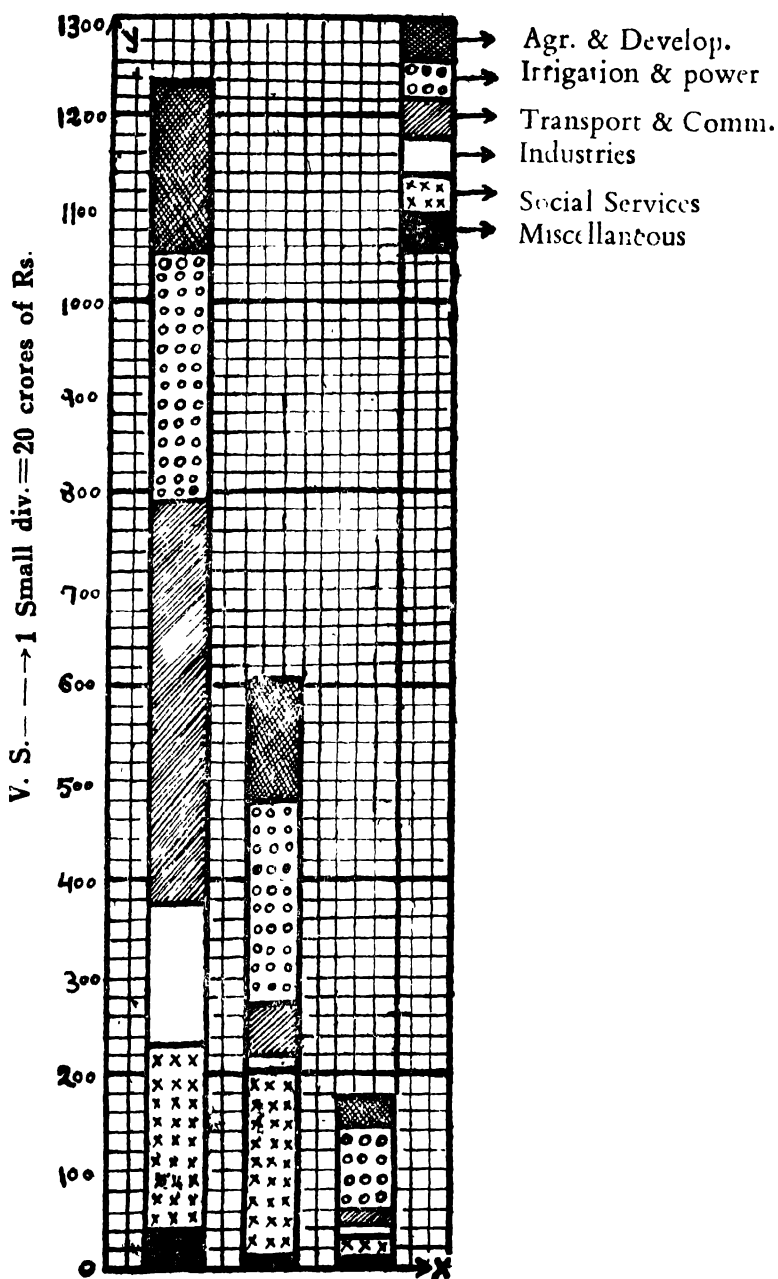
<i>Subject</i>	<i>Central</i>	<i>Type A States</i>	<i>Type B States</i>
Agr. & development	186·3	127·3	37·6
Irrigation & power	265·9	206·1	81·5
Transport & Communication	409·5	56·5	17·4
Industries	146·7	17·9	7·1
Social Services	191·4	192·3	28·9
Miscellaneous	40·7	10·0	7·0
Totals	1240·5	610·1	173·2

Solution:—

The data can be represented by means of a sub-divided bar diagram. Three bars will be constructed and their heights would be in proportion of the total expenditure of the three types of governments respectively. Further each bar will be sub-divided into six parts and the height of each part will be proportional to the expenditure incurred on it.

Graphs & Diagrams Diagram No. 4

59



C. GOV. G. of T. A. G. of T. B.

The diagram provides the information about the total expenditure of the central, type A and type B governments. This also gives the distribution of the total expenditure of three types of the governments with respect to the items of expenditure (Transportation, Agriculture etc).

Example No. (16)

Represent the data given below by a suitable diagram. The table gives the birth rates and the death rates of the six countries of the world during the year 1937.

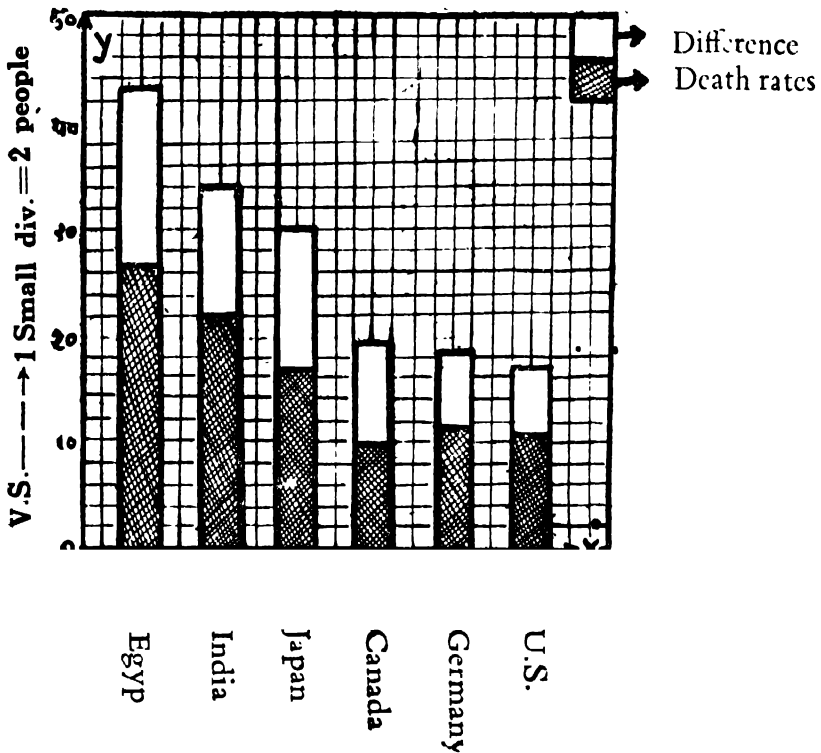
<i>Name of the Country</i>	<i>Birth rate</i>	<i>Death rate</i>
Egypt	43·5	27·2
Canada	19·8	10·2
United States	17·0	11·2
India	34·5	22·4
Japan	30·6	17·0
Germany	18·8	11·7

Solution:—

These birth and death rates can be represented by sub-divided bar diagrams. Six bars are constructed with their heights proportional to the birth rates. Further the bars are sub-divided into 2 parts, the lower one shows the death rates and the remaining upper portion shows the difference between the birth and the death rates.

Diagram No. 5

Showing the birth & death rates
of
six countries in 1937



H.S. ———→ 2 Small divs. = width of each bar
 " = Spacing between two consecutive bars

Percentage sub-divided bar-diagram:—

In this diagram, the total values are taken equal to 100 and the component parts are expressed in percentages. As the length of each bar in this diagram is the same, so it cannot give the comparison between the absolute magnitudes of the components. But the relative changes in the component parts are satisfactorily compared.

Example (17):—

With the help of a suitable diagram, show the absolute as well as the relative changes in the students population of the college A and B in the different faculties in 1947.

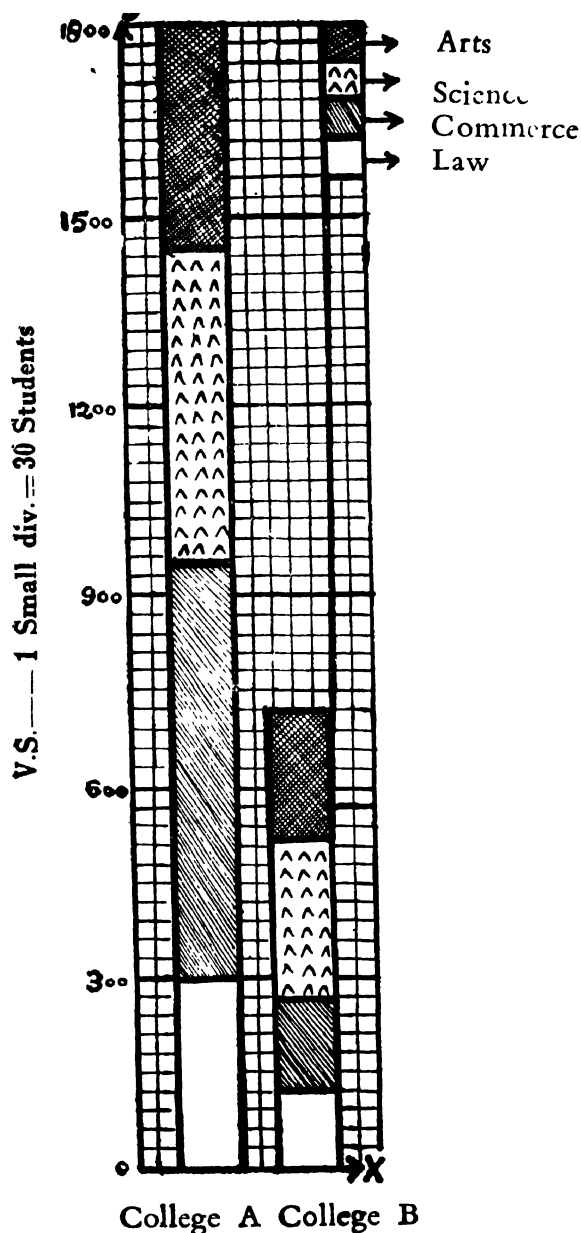
<i>Faculty</i>	<i>College A</i>	<i>College B</i>
Arts	350	290
Science	500	250
Commerce	650	150
Law	300	150
Totals	1800	720

Solution:—

(a) For the comparison of the absolute changes in the students population, the sub-divided bar diagram will be a suitable diagram.

Diagram No. 6 (a)

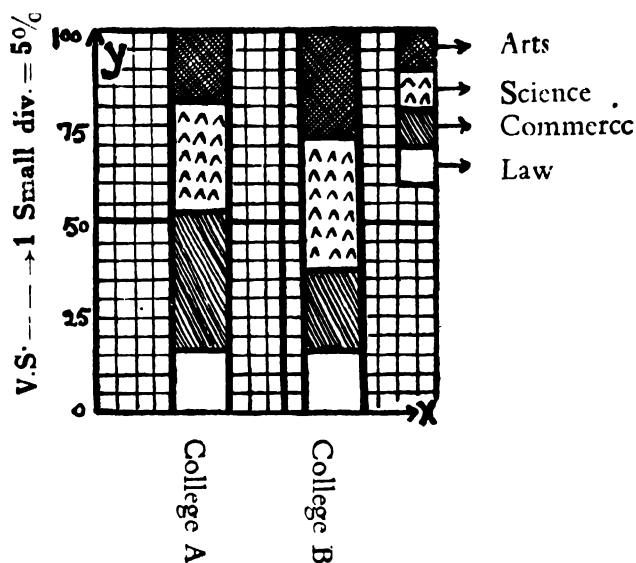
howing the absolute changes in Students population in 1947



(b) The relative changes in the component parts can be shown by the percentage sub-divided bar diagram. For this purpose, first we change the absolute magnitudes into the percentages.

Faculty	College (A)		College (B)	
	absolute magnitude	Relative magnitude in %	absolute magnitude	Relative magnitude in %
Arts	350	$\frac{350}{1800} \times 100 = 19.44$	200	$\frac{200}{720} \times 100 = 27.77$
Science	500	$\frac{500}{1800} \times 100 = 27.77$	250	$\frac{250}{720} \times 100 = 34.72$
Commerce	650	$\frac{650}{1800} \times 100 = 36.13$	150	$\frac{150}{720} \times 100 = 20.85$
Law	300	$\frac{300}{1700} \times 100 = 16.66$	120	$\frac{120}{720} \times 100 = 16.66$
Totals	1800	100	720	100

Diagram No. 6 (b)
Showing the Relative Change in
Students population in 1947 .



H. S. — — — — — 3 Small divs. = width of each bar
 4 " " = spacing between two bars

Example (18)

For the following data, show by a suitable diagram the comparison between the relative and the absolute changes.

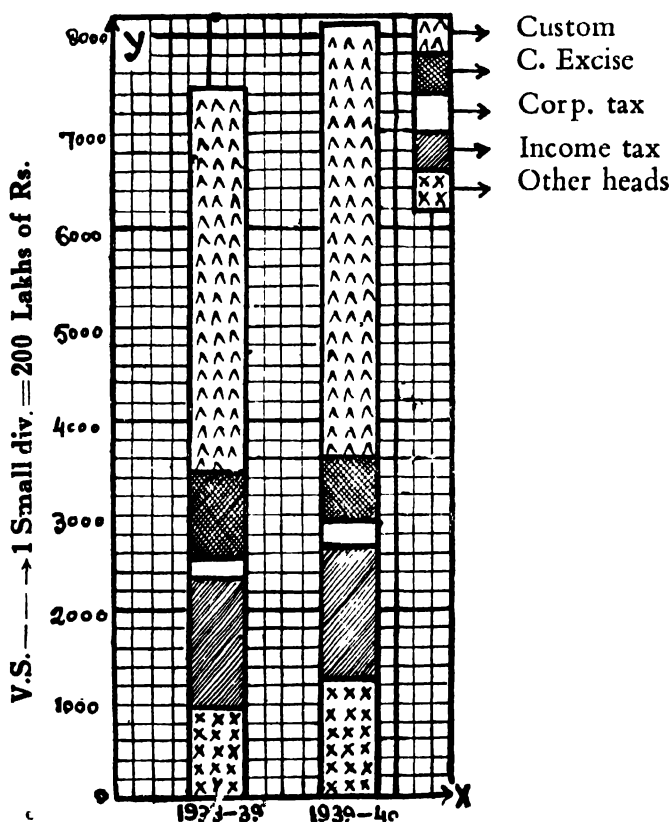
Principal heads of Revenue	Year 1938—39 (Lakh of Rs.)	Year 1939—40 (Lakh of Rs.)
Custom	4050	4588
Central Excise	868	652
Corporation tax	204	238
Incometax	1374	1420
Other heads	974	1262

Solution:—

(a) For the comparison of the absolute changes, the subdivided bar diagram will be a suitable diagram.

Digram No. 7(a)

Showing absolute changes in amounts incurred on various heads in 1938-39 & 1939-40

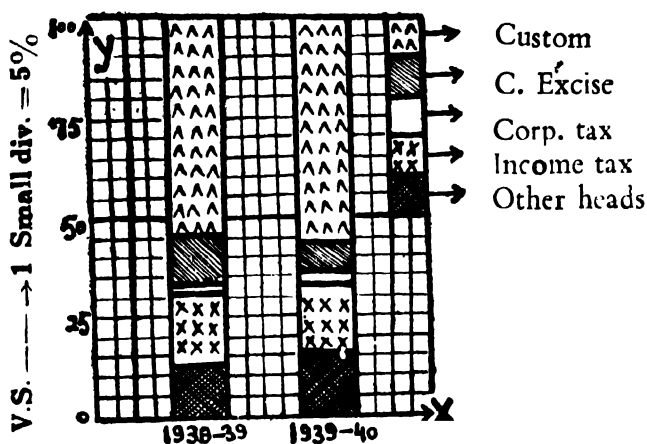


H. S. — — — — — 3 Small divs. = width of each bar = 1 year
 4 " " = spacing between two bars

(b) For the relative changes in the component parts, the percentage sub-divided bar diagram will be constructed. For this purpose, we change the absolute magnitudes into the percentages as shown below—

Heads of Revenue	Year 1938—39		Year 1939—40	
	Absolute magnitude	Relative magnitude in %	Absolute magnitude	Relative magnitude in %
Custom	4050	54.3	4588	56.4
Central Excise	868	11.6	652	8.0
Corporation tax	204	2.7	238	2.9
Income tax	1374	18.4	1420	17.5
Other heads	974	13.0	1262	15.2
Totals	7470	100	8160	100

Diagram No 7 (b)
Showing Relative changes in amounts
incurred on various heads in
1938—39 & 1939—40



H. S. ———> 3 Small divs. = width of each bar = 1 year
 4 " " = Spacing between two bars

Bilateral Bars: In this case, the bars are drawn above or below the base line (in the case of vertical bars) and to the left and right perpendicular to the base line (in the case of horizontal bars). The bars above the base line or to the right of the base line are used for the +ve quantity and the bars below the base line or to the left of the base line are used for —ve quantity in the data at hand.

Example (19):—

The following table gives the number of houses in ten small towns of India during the census of 1941 and 1951.

Town :	1	2	3	4	5	6	7	8	9	10
No. of houses										
in 1941:	200	300	400	480	520	300	400	280	750	570
No. of houses										
in 1951:	250	340	420	495	530	290	380	250	710	520

Represent the increase or decrease in the number of houses in 1951 in comparison to the census of 1941.

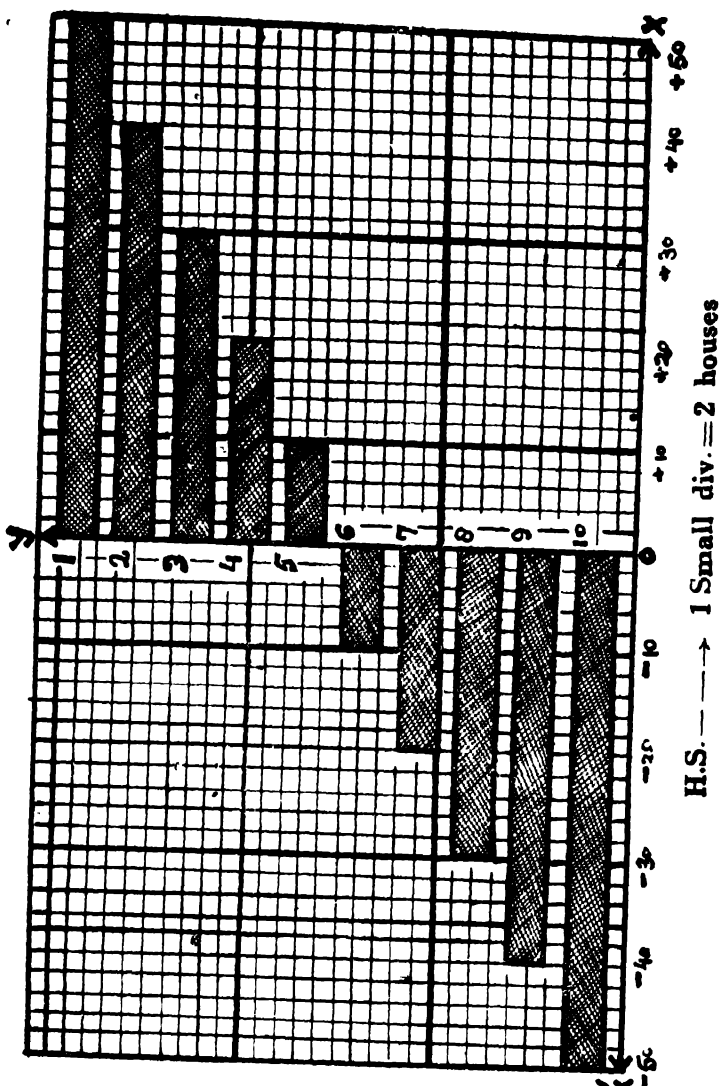
Solution —

In some cases, there is an increase while in other there is a decrease. The arrangement is made such that the towns which show an increase are written first and the towns which show decrease are written later on. Within these graphs, the towns have been arranged according to the magnitudes of the changes (increase or decrease).

Town	No. of houses in		Difference
	1941	1951	
1	200	250	+50
2	300	340	+40
3	400	420	+20
4	480	495	+15
5	520	530	+10
6	300	290	-10
7	400	380	-20
8	280	250	-30
9	750	710	-40
10	570	520	-50

In the diagram, the +ve changes (increase in the no. of houses) are shown by the bars drawn to the right of the base line and those showing the -ve changes (decrease in the no. of houses) are drawn to the left of the base line.

Diagram No. 8



Example No (20):-

Represent the following data by a suitable diagram showing the differences between the proceeds & costs.

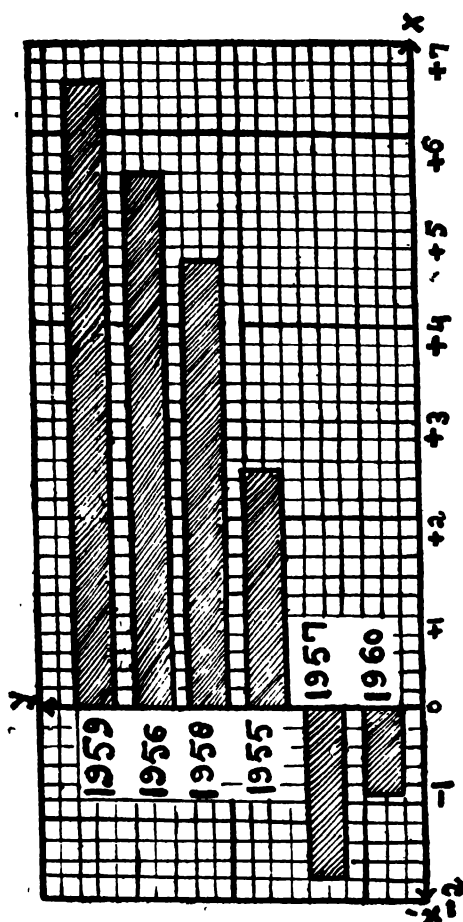
Year	Total Proceeds (in thousands of Rs.)	Total costs (in thousands of Rs)
1955	22·0	19·5
1956	27·3	21·7
1957	28·2	30·0
1958	30·3	25·6
1959	32·7	26·1
1960	33·3	34·2

Solution:—

First of all we take the differences of the total proceeds & the total costs and then +ve differences are written first and —ve differences after them. Within these groups, they have been arranged according to the magnitudes of the changes (increase or decrease).

<i>Year</i>	:	1959	1956	1958	1955	1957	1960
<i>Difference:</i>		+6·6	+5·6	+4·7	+2·5	—·90	—1·8

Diagram No. 9
Differences between Proceeds & Costs
from 1955-60



H. S. ———→ 1 Small div. = 200 Rs.

(2) Two Dimensional Diagrams : (Area Diagrams)

In these diagrams, the areas are proportional to the magnitudes of the data. The common diagram of this type are rectangles, pie and square diagrams. The square and pie diagrams serve the same purpose but the pie diagrams are the easiest to draw and they can be made accurately. Hence the pie diagrams are commonly used in place of square diagrams.

(a) **Pie/circle diagrams** : When the differences between any two quantities to be compared are large, bar diagrams can not be used, as one of them will be extremely small.

and the other large. In such cases, the squares or Pie diagrams are used. For drawing the circles, the square roots of the quantities are taken as the radii of the circles and so the areas of the circles are proportional to the magnitudes of the data.

(b) Sub-divided Pie diagrams : When we want to compare the totals as well as their components with one another, we may use sub-divided Pie diagrams. The total value is equaled to $2\pi=360^\circ$ and then component parts are expressed in terms of angles. Having determined the angles corresponding to different components, the circle is divided into different sectors on the various angles pre-determined.

Example No (21):—

Represent the following data by a suitable diagram—

**Countries Production of cane sugar in 1938-39
in Quintals 0000's omitted**

1. India	2750
2. Java	1550
3. Hawaii	835
4. Columbia	51

Solution:—

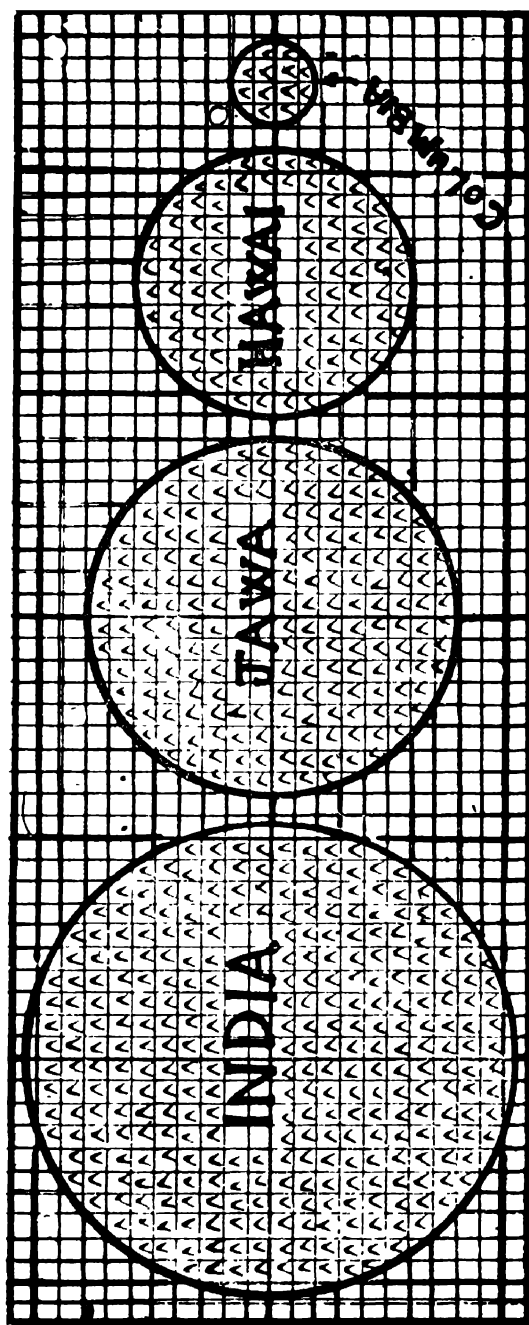
Here the difference between the quantities 51 & 2750 is very large. Thus Pie diagram will be a suitable diagram to represent the data. Now we construct the following table—

Countries	Quintals (1) 0000' S omitted	Square roots (2)	[.] Radii in inches
1. India	2750	52.44	1.05
2. Java	1550	39.37	0.79
3. Hawaii	835	28.90	0.58
4. Columbia	51	7.14	0.14

The column (2) gives the square roots of the figures written in column (1) and the column (3) contains the numbers which are obtained by dividing the numbers of the column (2) by 50 which are the radii in inches. The circles with these radii are arranged in the descending order. For finding the scale, we calculate the area of the 1st circle = $\frac{22}{7} \times (1.05)^2 = 3.14 \times 1.1025$ sq. inch.

Thus 3.14×1.025 sq. inches = 2750.0000 quintals
 \therefore 1 sq. inch = 796,5000 quintals.

Diagram No. 10
Production of Cane Sugar in Certain Countries during 1938—39



Scale: 1 sq. inch=796,500 quintals.

Example 22:—

The allocations for Madras & U.P. States under the second 5 years plan are as given below—

Heads	Madras State (lakhs of Rs.)	U.P. State (lakhs of Rs.)
1. Agr. & Cumunity Development	3535.5	6763.9
2. Irrigation & Power	7125.0	8042.5
3. Industry & mining	1520.0	1643.4
4. Transport	807.5	1723.2
5. Social Services	4065.9	6863.8
6. Miscellaneous	252.1	272.8

Draw a sub-divided Pie diagram to compare the cost of development under each head in the two states?

Solution:—

The solution for the problem is given in the tabular form as shown on the next page.

Heads	Madras State		U. P. State		Square roots	Radii in inches (dividing by 100)
	Cost	angle	Cost	angle		
1. Agr. & community development	3535.5	73.5	6763.9	96.04	$\sqrt{17.06} = 131.5$	1.3
2. Irrigation & Power	71 5.0	148.2	8042.5	114.20	$\sqrt{25309.6} = 159.09$	1.6
3. Industry & mining	1520.0	31.6	1643.4	23.34		
4. Transport	807.5	16.8	1723.2	24.47		
5. Social services	4065.9	84.6	6863.8	97.47		
6. Miscellaneous	252.1	5.3	272.8	4.48		
Totals	17306.0	360°	25309.6	360°		

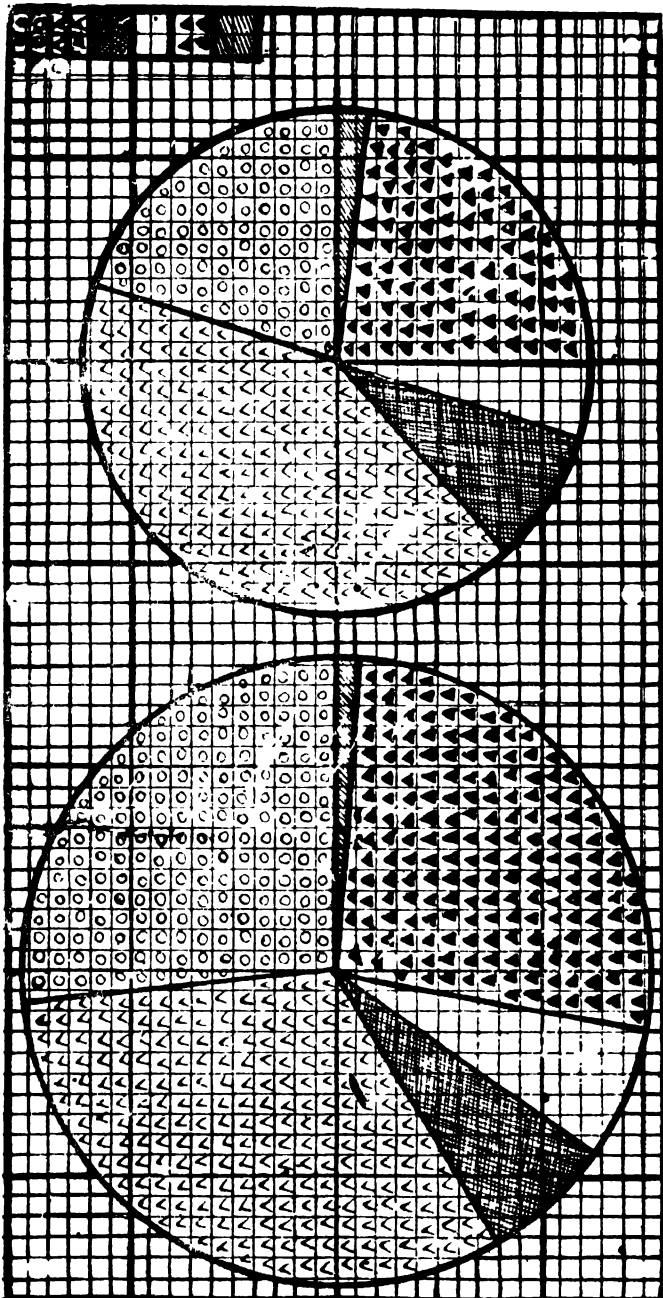
Steps to construct the diagram : We take the square roots of 17306.0 and 25309.6 which are 131.5 and 159.09 and then divide these square roots by any number (say 100 here) to obtain the radii of the circles. Then the sectors on various angles are cut from these circles according to the angles shown in the table. In this way, the areas represented by different sectors on a circle are proportional to the magnitudes of the components contained for a state as the main heads. Since $3.14 \times (1.3)^2 = 5.31$ equivalents to 17306 lakhs of Rs, so Scale is 1 sq. inch = 3259.1 Lakhs of Rs.

Graphs & Diagrams

Diagram No. 11

U.P.

MADRAS



Agr. & C.
Irrig. & P.
Ind. & M.
Transport
Soc. S.
Misc.

Example (23):—

Represent the data of the following table by means of a suitable sub-divided Pie diagram—

Clearing house Statistics in 1940-41 and 1947-48 in certain cities—

City	Total amount (Rs.)	
	1940-41	1947-48
Bombay	80,232	255,264
Calcutta	100,853	259,996
Delhi	2,853	12,646
Kanpur	1,920	10,983
Karanchi	4,676	27,481
Lahore	1,633	4,954
Madras	10,865	34,794
Others	4,228	51,896
Totals	2,07,260	658,014

Solution: —

The steps in the construction of Pie diagram are given below—

(i) First of all, calculate the square-roots of the total amount of clearing house returns for the two years i.e. calculate $\sqrt{207260}$ and $\sqrt{658014}$ which are 455 & 811 respectively.

(ii) Divide these quantities (455, 811) by a suitable number to obtain the radii of the circles.

(iii) Then calculate the angles corresponding to the amounts for each city in the separate sessions. The calculations are done in the following tabular manner—

Cities	Year 1940—41 Angles calculated (degree)	Year 1947—48 Angles calculated (degree)
Bombay	$\frac{80232 \times 360}{207260} = 139.4$	$\frac{255264 \times 360}{658014} = 139.7$
Calcutta	$\frac{100853 \times 360}{207260} = 175.2$	$\frac{259996 \times 360}{658014} = 142.3$
Delhi	$\frac{2853 \times 360}{207260} = 5.0$	$\frac{12646 \times 360}{658014} = 6.9$
Kanpur	$\frac{1920 \times 360}{7260} = 3.3$	$\frac{10983 \times 360}{658014} = 6.0$
Karanchi	$\frac{4676 \times 360}{207260} = 8.1$	$\frac{27481 \times 360}{658014} = 15.0$
Lahore	$\frac{1633 \times 360}{207260} = 2.8$	$\frac{4954 \times 360}{658014} = 2.7$
Madras	$\frac{10865 \times 360}{207260} = 18.9$	$\frac{34794 \times 360}{658014} = 19.0$
Others	$\frac{4228 \times 360}{207260} = 7.3$	$\frac{51896 \times 360}{658014} = 28.4$
Totals	360	360

Finally we divide 455 & 811 by a number say 400 to get 1.14 and 2.05 as the approximate radii in inches and the sectors on various angles are cut from these circles. The area enclosed within each sector represents the magnitude of the amount returns in the specified years for different cities.

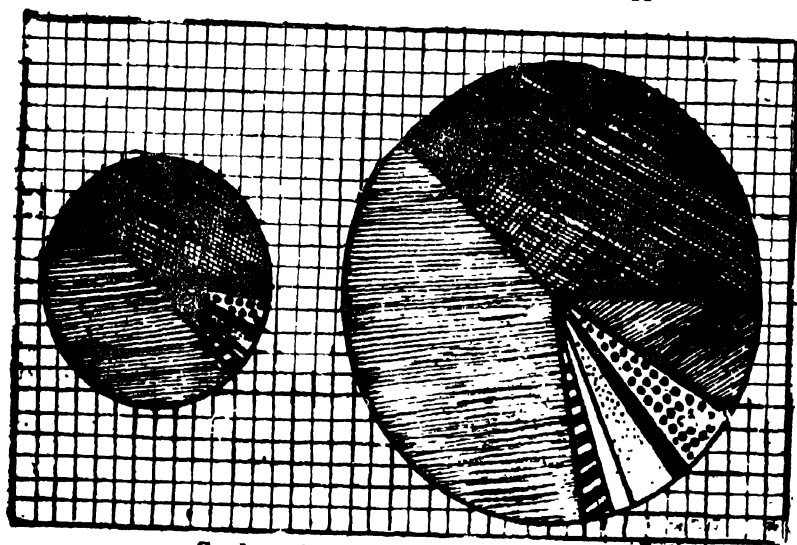
Scale : 1 sq. inch = 52389 Rs.

Diagram No. 12

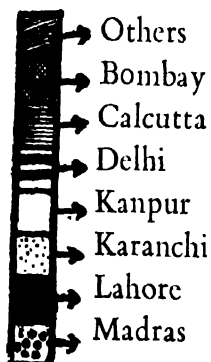
'representing clearing house statistics'

Year : 1940-41

Year : 1947-48



Scale : 1 sq. inch = 52389 Rs.



Example 24—

The following table gives the total outlay on rural development proposed in the first Five Years Plan and its break down into major items. Give a suitable diagrammatic representation of the data.

(M.Sc Ag.Agra, 1965)

Item	Amount (in crores of rupees)
Agr. & community development	360.43
Irrigation	167.97
Irrigation & Power (multipurpose projects)	265.90
Power	127.54
Transport & communications	497.10
Industry	173.04
Social services	339.81
Rehabilitation	85.00
Miscellaneous	51.99
Total	2,068.78

Solution:—

The suitable diagram for representing the above data is sub-divided Pie diagram. The total area of the circle will represent the total amount i.e. 2068.78 crores of rupees and its components will represent the various items of expenditure under the plan.

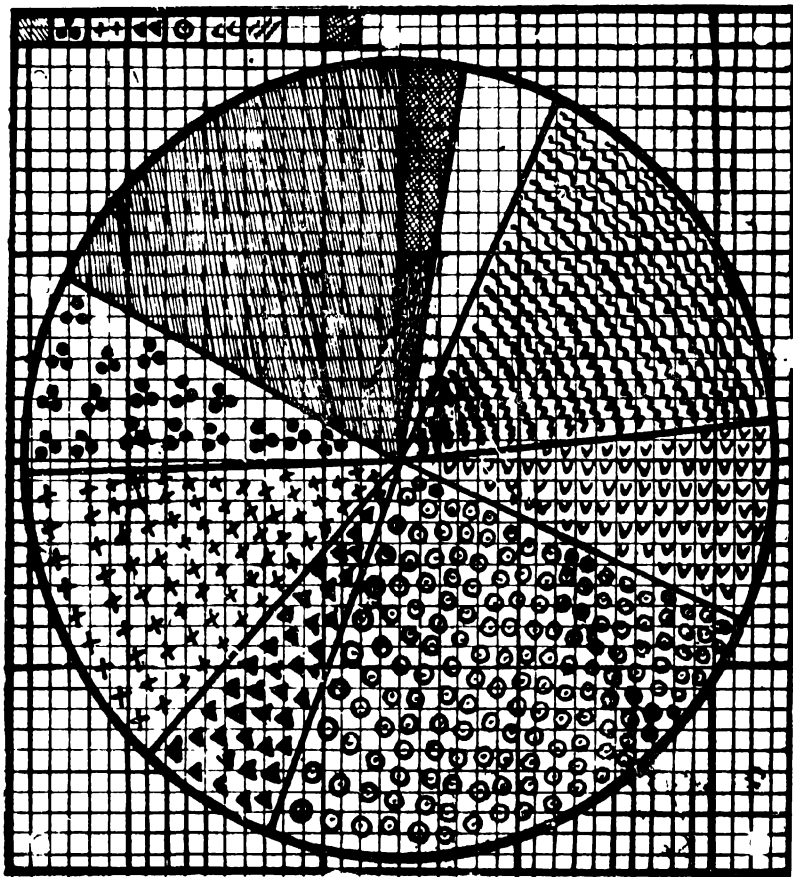
Item	Amount (in crores of Rupees)	Angle
Agr. & com. development	360.43	62.7°
Irrigation & power.... ..	167.97	29.2°
Irrigation... ..	265.90	46.3°
Power.... ..	127.54	22.2°
Transport & communication....	497.10	86.5°
Industry	137.04	30.1°
Social services.....	339.81	59.1°
Rehabilitation	85.00	14.8°
Miscellaneous.....	51.99	9.0°
Totals	2068.78	$359.9^\circ \approx 360^\circ$

Let the radius of the circle be $\frac{2}{7}$ ", then $\pi r^2 = \text{Area}$
 or $\frac{2}{7}^2 \times 4 \text{ sq. inches} = 2068.78 \text{ crores of rupees.}$

Scale : 1 sq. inch = 164.56 crores of rupees.

Diagram No. 13
Showing the total outlay on Rural development

- Agr. & com. development
- Irrigation & Power
- Irrigation
- Power
- Transport & Communication
- Industry
- Social Services
- Rehabilitation
- Miscellaneous



Scale : 1 sq. inch = 164.56 crores of Rs.

Rectangles : The rectangles are the two dimensional diagrams and are used when the two magnitudes which are related to a third one are to be represented on the same diagram. For example, the average yield of wheat multiplied by the total acreage under wheat represents the total production of wheat. To represent it diagrammatically, one side of the rectangle is taken proportional to the average-yield and the other is proportional to the total acreage; the area of the rectangle gives the total production of wheat. Sub-divided rectangles are used when one of the magnitudes is divided into several components.

Example No (25)

An analysis of the monthly wages paid to workers in the two firms A & B gives the following results—

	Firm A	Firm B
No. of wage earners	80	100
Average monthly wage (Rs.) :	52.5	47.5

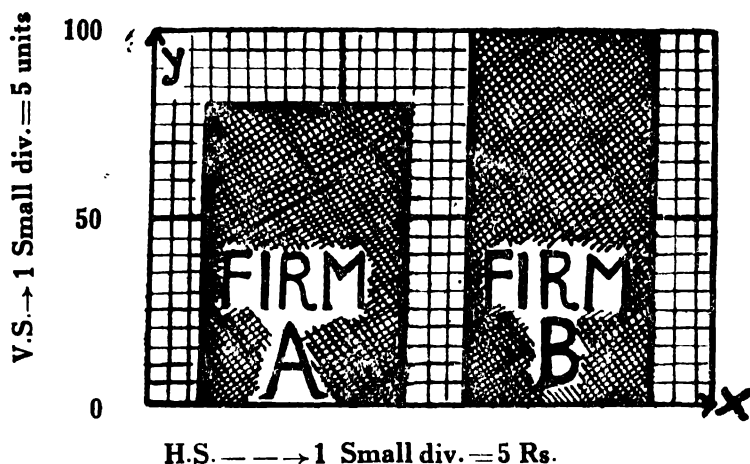
Represent the above facts diagrammatically ?

Solution:—

Here the average wage multiplied by the total number of workers gives the daily paid roll. Hence the most suitable diagram for representing the above data is to construct the two rectangles, one for each firm. One side of the rectangles will be proportional to the average wage and the other will be proportional to the total no. of workers employed in a firm.

Diagram No. 14

Showing the average monthly wage and number of wage earners



The family budget data, where we compare the absolute totals as well as the percentages of the various items to the total, is very suitably represented by the rectangles. In this case, the height of each rectangle is kept 100 and the width is made proportional to the total magnitude.

Example No. 26

The following table gives the details of the monthly expenditure of three families. Represent them by a suitable diagram--

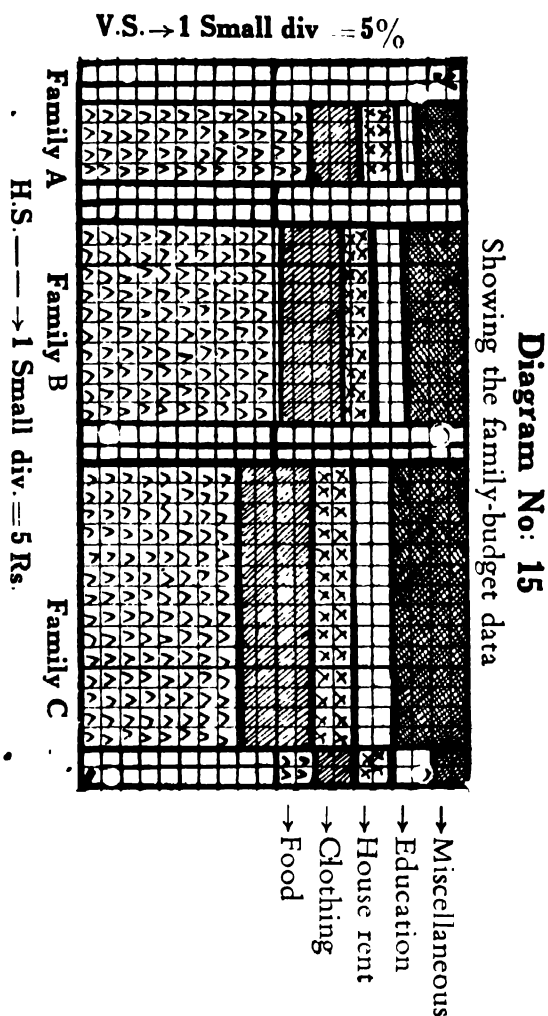
Item of the expenditure	Family A	Family B	Family C
	Rs. An.	Rs. An.	Rs. An.
Food	12—0	25—0	28—0
Clothing	2—8	8—0	14—0
House rent	2—0	4—0	7—
Education	1—0	5—0	7—0
Miscellaneous	2—8	8—0	14—0
Totals	20—0	50—0	70—0

Solution: —

First we prepare the following table:

Item of expenditure	Family A		Family B		Family C	
	Actual expenses	%	Actual expenses	%	Actual expenses	%
Food	12.00	60.00	25.00	50.00	28	40.00
Clothing	2.00	12.50	8.00	16.00	14	20.00
House rent	2.00	10.00	4.00	8.00	7	10.00
Education	1.00	5.00	5.00	10.00	7	10.00
Miscellaneous	2.50	12.50	8.00	16.00	14	20.00
Total	20.00	100	50	100	70	100

The heights of the three rectangles will be taken as 100 for each and their width in the ratio 20:50:70. Each rectangle will be further divided into component parts according to the figures in the percentage column.



The above diagram represents—

(i) The actual expenses on each item by the area of the corresponding component part of the rectangle.

(2) The % of the item to the total by the height of the corresponding component part of the rectangle.

(3) The total expenses for the family by the whole area of the rectangle.

Example (16):—

Represent the following data by a suitable diagram —

Year	Average yield Md/acre	Total yield	Area under crop		Total (acre age)
			Irrigated	Non-Irrigated	
1940- 41	9.5	427.5	22	23	45
1941- 42	8.4	361.2	23	20	43

Solution:—

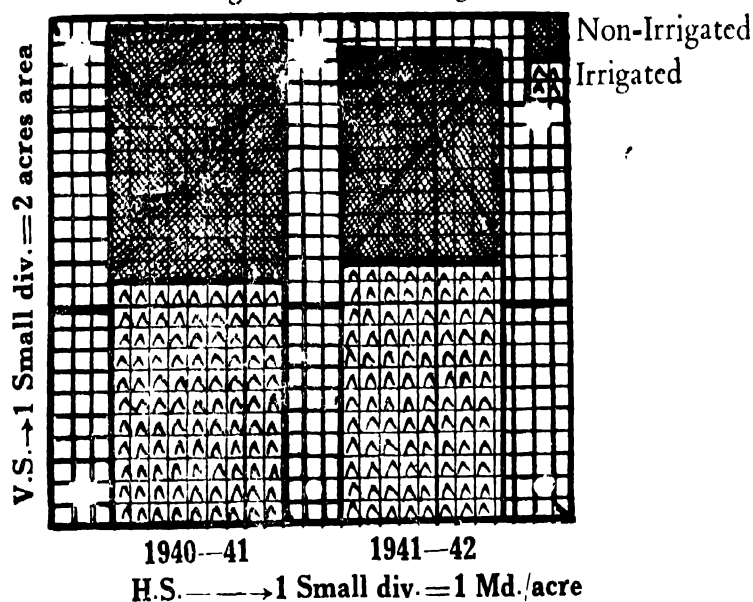
Here we note the relation—

Average yield \times total area = total yield.

So the rectangles are the suitable diagrams to represent the given data. Here we construct the two rectangles on the same horizontal axis taking their width in the ratio 9.5:8.4 and heights in the ratio 45:43. Each rectangle will be further divided into two components according to the irrigated and non-irrigated areas.

Diagram No. 16

Showing the average yield md/acre & area under crop
irrigated and non-irrigated



Three dimensional Diagrams : When the ratio between the two quantities to be compared is so large that the two dimensional diagrams are not suitable for representing them, we may represent them by three dimensional diagrams. The volumes of the diagrams represent the magnitudes of the quantities to be compared. In this type of diagrams the spheres, cubes, and the prisms are common but cubes are generally drawn in practice.

To calculate the arithmetic means of the milk yield, we prepare the following table—

Milk yield (1) X (in Kgm.)	No. of cows (2) f	(3) fx
11	6	66
14	9	126
17	12	204
21	16	336
25	21	525
30	9	270
32	7	224
Totals	80 = Σf	1751 = Σfx

In column (1), we put the variate values and in column (2) their respective frequencies against them. In column (3) we put the product of (1) & (2). For example, the first value 66 of the colⁿ. (3) is the product of 11 & 6 which are the 1st. figures in the colⁿ. (1) & (2) respectively. Similarly, the other figures of colⁿ. (3) are attained. Thus, we get—

$$\bar{X} = \frac{\Sigma fx}{\Sigma f} = \frac{\text{sum of col}^n. (3)}{\text{sum of col}^n. (2)} = \frac{1751}{80} = 21.8875 \text{ kgms.}$$

This formula can be applied in a grouped frequency distribution after the classes have been replaced by their respective mid-values. The procedure of calculation will be clear from the following Exp.—

definite which can be determined mathematically

(2) It should be based on all observations

(3) Its calculation should not be lengthy and tedious.

(4) It should be least affected by sampling fluctuations.

(5) It must be capable of algebraic treatments i.e. if the averages of the component series are known then the average of the whole series should be expressible in terms of the averages of the component series.

Relative merits & Demerits of different averages :

(1) **Mean** (Arithmetic average)

Advantages:—

(1) It is readily understood and well defined.

(2) It is based on all observations.

(3) It is easy to calculate.

(4) In most of the cases, it is not affected by the sampling fluctuations.

(5) It is capable of algebraic treatments.

(6) It gives weights to all items which are directly proportional to their sizes.

Disadvantages:—

(1) It some times gives values which may not be physically possible e.g. the average number of eggs laid by a hen as 18.5 per month.

(2) It gives undue weights to the extreme items.

(3) It can not be calculated in the cases where the extreme ends are open.

(2) **Mode :**

Advantages:—

(1) In most of the cases, it is easily found.

(2) It can be found out from the graph merely by an eye inspection.

(3) It can be found out for the distributions where the ends are open.

(4) It is the type that to the ordinary mind, seems to be the best to represent the group.

Disadvantages:—

- (1) No weights are given to the extreme items.
- (2) A clearly defined mode does not always exist
- (3) It is not capable of algebraic-treatments

(3) Median :

Advantages:—

- (1) It is well defined and can be calculated easily.
- (2) It does not give undue weights to the extreme items.
- (3) It is possible to calculate even in the cases where the intervals are open.

Exp. No. 2

What is an arithmetic average ? What are its properties ?
How would you calculate the Arithmetic Average ?

Sol: —

The arithmetic mean of a series of individuals is obtained by adding up all the values and then dividing the total by the number of individuals.

For example, if the heights of 10 randomly chosen plants are—52", 55", 57", 61", 64", 65", 67", 68", 70", & 71", then the arithmetic mean or simply the mean of this sample of 10 plants will be the sum (52 + 55 + + 71 = 630) divided by 10. i.e. mean = $\frac{630}{10} = 63"$

In general, if x_1, x_2, \dots, x_n are the values of n measurements, the arithmetic mean \bar{X} of the X_s is defined by the relation—

$$\bar{X} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\Sigma x}{n} \dots \dots \dots (i)$$

where x_1, x_2, \dots, x_n are the given values

n = number of items

Σ = the sum of

\bar{X} = Arithmetic mean of X_s

Now, the formula (i) will be read as "The arithmetic mean of x_s is the sum of x_s divided by their number."

Exp. No. 5

If x_1, x_2, \dots, x_n be the 'n' values of any measurement with their respective frequencies f_1, f_2, \dots, f_n and mean \bar{X} , then show that

$$\bar{X} = A + \frac{\sum fd}{n},$$

where A is the assumed mean and

d is the deviation of x from A

i.e. $d = (x - A)$

Sol —

$$\begin{aligned} \text{We define } \bar{X} &= \frac{\sum fx}{n} \\ &= \frac{\sum f(x - A + A)}{n} \\ &= \frac{\sum f(x - A)}{n} + \frac{A \sum f}{n} \\ &= \frac{\sum f(x - A)}{n} + A \cdot \frac{n}{n} \\ \therefore \bar{X} &= A + \frac{\sum f(x - A)}{n} \dots \dots \dots (i) \\ \text{or } \bar{X} &= A + \frac{\sum fd}{n} \end{aligned}$$

In future, for the computation of mean, we shall apply the above (i) formula which reduces the bulk of calculations and hence called the *short cut formula for mean*.

Exp. No. 6

Using the short cut method, compute the mean of the following data—

(a) No. of branches/

plant (x): 5 10 15 20 25 30 35 40 45 50

No. of plants (f): 20 43 75 67 72 45 39 9 8 6

(b) Age (years) 0—10, 10—20, 20—30, 30—50, 50—80

No. of persons : 61, 49, 40, 60, 23

Chapter III

Measures of Central tendency (averages) and dispersion

The histogram or frequency curve gives the general idea of the distribution of the variate and hence the frequency graph can be used to study and compare the given distributions. But the study through the graphs depends upon the accuracy and skill of the eye, which is an uncertain factor. Thus the study and comparisons by it may not be very reliable. Therefore it is necessary to know certain features of the distribution, which give an idea of the distribution and can be determined arithmetically. Two such features of the distribution are its average and dispersion (variation).

Average : An average is the value of the variate which claims to represent the distribution. Some of the variate values will be above this value and others below it and so it is known as a measure of the central tendency.

There are three averages (i) Mean (ii) Mode and (iii) Median.

Exp. No. 1

What are the properties of an ideal average. In what circumstances, would you consider the mean, mode and the median, the most suitable statistics to describe the central tendency of the distribution ?

Sol:—

An ideal average should have the following properties:—

(1) It should be rigidly defined and its value should be

Exp. No. 4

Given the following distribution, calculate the mean:

Height of plants (in cms.)	No. of plants	Height of plants (in cms.)	No. of plants
12.5—17.5	2	37.5—42.5	4
17.5—22.5	22	42.5—47.5	6
22.5—27.5	19	47.5—52.5	1
27.5—32.5	14	52.5—57.5	1
32.5—37.5	3		

Solution:—

The mean is given by the formula $\bar{X} = \frac{\sum fx}{n}$, where the computations of $\sum fx = n$ are given below in the tabular form—

(1) x	(2) f	(3) fx
15	2	30
20	22	440
25	19	475
30	14	420
35	3	105
40	4	160
45	6	270
50	1	50
55	1	55
Totals	72 = $\sum f$	2005 = $\sum fx$

In the colⁿ. (1) we have put the mid-values of the classes and in (2) the class frequencies respectively. Thus, the product of (1) & (2) are put in the colⁿ. (3)

Therefore, Σfx is the sum of all figures in the colⁿ. (3) and $\Sigma f = n$, the sum of all figures in the col. (2). Hence we have

$$\bar{X} = \frac{\Sigma fx}{n} = \frac{2005}{72} = 27.85 \text{ centimetres.}$$

For a short cut method for computing the arithmetic-mean, the formula is—

$$\bar{X} = A + \frac{\Sigma fd}{n}$$

where A is the assumed mean and

$d = (x - A)$ the deviation of the variate from the assumed mean.

Properties of the Arithmetic Mean

(1) The sum of the deviations of the values of x_s from their mean \bar{X} is always zero i.e. if x_1, x_2, \dots, x_n be the variate values and \bar{X} be their mean, then $(x_1 - \bar{X}), (x_2 - \bar{X}), \dots, (x_n - \bar{X})$ etc. are their deviations from the mean and the sum of deviations is zero. i.e. $\Sigma (x - \bar{X}) = 0$

(2) If $\bar{X}_1, \bar{X}_2, \dots, \bar{X}_k$ be the arithmetic means of k distributions with respective frequencies n_1, n_2, \dots, n_k , then the mean \bar{X} of the whole distribution is given by—

$$\bar{X} = \frac{n_1 \bar{X}_1 + \dots + n_k \bar{X}_k}{n_1 + \dots + n_k} = \frac{\Sigma n \bar{X}}{\Sigma n}$$

(3) If x_1, x_2, \dots, x_n be the variate values with their mean \bar{X} and a variate $y = ax + b$ is obtained, then $\bar{Y} = a\bar{X} + b$, where a & b are constants.

Exp . No. 3

The yield of rices in 12 equal plots of a village are given as follows (in maunds) —

68, 80, 94, 104, 120, 114
125, 130, 130, 140, 141, 145, find the mean yield ?

Solution:—

If the yields are denoted by x_s , then

$$\bar{X} = \frac{\sum x}{n} = \frac{68+80+\dots+145}{12} = \frac{1392}{12} = 116 \text{ maunds/plot.}$$

If the value x_1 occurs f_1 times, x_2 occurs f_2 times, and so on, then

$$\bar{X} = \frac{f_1 x_1 + f_2 x_2 + \dots + f_n x_n}{f_1 + f_2 + \dots + f_n} = \frac{\sum f x}{\sum f} = \frac{\sum f x}{n} \dots\dots\dots (ii)$$

as $\sum f = n$

where \bar{X} = Arithmetic average and x_1, x_2, \dots, x_n are the given values of the variate x_1 .

f_1, f_2, \dots, f_n are the respective frequencies of x_1, x_2, \dots, x_n .

$$n = f_1 + f_2 + \dots + f_n = \sum f$$

To make the procedure clear, let us take the yield of milk by 80 cows of a dairy farm on a certain day (in kilograms) —

Milk-yield : 11 14 17 21 25 30 32

No. of cows 6 9 12 16 21 9 7.

Solution:—

(a) To calculate the mean of this data, first we perform the following table—

$$A=25 \checkmark$$

(1) x	(2) f ✓	(3) d=x-A	(4) fd
5 ✓	20	—20	—400
10	43	—15	—645
15	75	—10	—750
20	67	—5	—335
25 ✓	72	0	0
30	45	5	225
35	39	10	390
40	9	15	135
45	8	20	160
50	6	25	150
Totals	384=Σf	—	—1070=Σfd

The assumed mean i.e A should be taken at such a value of x which divides the whole distribution approximately into two equal parts, so that some of the 'd' values be —ve and the others be +ve. Here, we have taken A=25 and colⁿ. (3) is obtained by subtracting A=25 from the figures of colⁿ. (1) as the first value in colⁿ. (3) is 5—25=—20 and so on, the others. The colⁿ. (4) is obtained as the respective product of (2) & (3). Then we compute

$$\bar{X}=A+\frac{\Sigma fd}{n}$$

$$=25+\frac{-1070}{384}=25-2.8=22.2$$

∴ $\bar{X}=22.2$ no. of branches/plant.

(b) To calculate the mean of this data, the following table is performed—

$$A=25$$

(1)	(2)	(3)	(4)
x	f	d = x - A	fd
5	61	-20	-1220
15	49	-10	-490
25	40	0	0
40	60	15	900
65	23	40	920
totals	233 = Σf		110 = Σfd

The colⁿ. (1) of this table has been obtained by replacing the classes by their respective mid-values and colⁿ. (3) is the result of subtracting $A=25$ (assumed mean) from the figures of colⁿ. (1). The colⁿ. (4) is the product of (2) & (3) colⁿs. with their respective figures. Then we compute the mean

$$\begin{aligned}\bar{X} &= A + \frac{\Sigma fd}{n} \\ &= 25 + \frac{110}{233} = 25 + 0.47\end{aligned}$$

$$\therefore \bar{X} = 25.47 \text{ years.}$$

Example No. 7

What is median ? Illustrate by an example, how would you calculate the median ?

Solution: —

Median : The median is defined as the value of a variate which divides the distribution into two equal parts. Thus, half of the values lie below the median and half above it.

Calculation of Median :

I—(a) Simple series of odd items :

. In a simple series, the variate does not repeat itself. In such a series, if the number of items is odd, the computation of the median is very easy.

First we arrange the values either in the ascending or in the descending order of their magnitudes and then find the size of $\left(\frac{n+1}{2}\right)^{\text{th}}$ item in this series. It will be our desired value of the median.

For illustration , consider the following example of 9 items with magnitudes 2, 5, 7, 3, 11, 9, 8, 4 & 6.

To find the median, the items are arranged in the ascending order of their magnitudes i. e.

2, 3, 4, 5, 6, 7, 8, 9, 11, then size of $\left(\frac{n+1}{2}\right)^{\text{th}}$ item gives the median. Therefore

$$\begin{aligned} \text{Md} &= \text{size of } \left(\frac{n+1}{2}\right)^{\text{th}} \text{ item} \\ &= \text{size of } \left(\frac{9+1}{2}\right)^{\text{th}} \text{ item} = \text{size of } 5^{\text{th}} \text{ item} \\ &= 6 \end{aligned}$$

$\therefore \text{Median} = 6$

I—(b) Simple series of even items :

In this case, the median is the simple arithmetic average of the size of $\left(\frac{n}{2}\right)^{\text{th}}$ item and $\left(\frac{n}{2} + 1\right)^{\text{th}}$ item when the values have been arranged either in the ascending or in the

descending order of their magnitudes. For example, take the case of finding the median from the data of eight (even number) values—

62, 65, 67, 64, 72, 53, 59, 55

The ascending arrangement is

53, 55, 59, 62, 64, 65, 67, 72. Then the size of

$$\left(\frac{n}{2}\right)^{\text{th}} \text{ item} = \text{size of } \left(\frac{8}{2}\right)^{\text{th}} \text{ i.e. } 4^{\text{th}} \text{ item} = 62$$

$$\& \text{ size of } \left(\frac{n}{2} + 1\right)^{\text{th}} \text{ item} = \text{size of } 5^{\text{th}} \text{ item} = 64$$

Hence, Md. = Simple arithmetic average of the size of

$$\left(\frac{n}{2}\right)^{\text{th}} \& \left(\frac{n}{2} + 1\right)^{\text{th}} \text{ items.}$$

$$= \frac{62 + 64}{2} = 63$$

$$\therefore \text{Median} = 63$$

Ungrouped data of discrete variate:

If a variate x_1 repeats f_1 times, x_2 repeats f_2 times and so on i.e. if we deal with a frequency distribution of the following type—

Variate value (x)	frequency (f)
x_1	f_1
x_2	f_2
x_3	f_3

\vdots
 x_n

\vdots
 f_n , then there are

following steps in the calculation of the median.

(i) Arrange the items either in ascending or descending order of their magnitudes, if they are not so. But usually in such a case the data is already arranged and a fresh arrangement is rarely required.

(ii) Compute the cumulative frequencies of the variate values.

Measures of Central tendency (averages) and dispersion 101

(iii) locate the median item which is $\left(\frac{n+1}{2}\right)^{\text{th}}$ where n is the total frequency.

(iv) Find the value of the median which is the size of the $\left(\frac{n+1}{2}\right)^{\text{th}}$ item.

Consider the example given below to compute the median—

$x : 1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6$

$f : 20 \quad 15 \quad 17 \quad 19 \quad 13 \quad 15$

For the computation of the median, we proceed as follows—

x	f	$c \cdot f$
1	20	20
2	15	35
3	17	52
4	19	71
5	13	84
6	15	99

Since the median is the size of the $\left(\frac{n+1}{2}\right)^{\text{th}}$ item, the median here, is the size of $\left(\frac{99+1}{2}\right)^{\text{th}} = 50^{\text{th}}$ item.

By inspecting the colⁿ. of cumulative frequencies we note that the 50th item falls opposite the variate value 3 and so the median here is 3.

In case, when the total of all frequencies is an even number, the median will be the mean of the size of $\left(\frac{n}{2}\right)^{\text{th}}$ and $\left(\frac{n}{2}+1\right)^{\text{th}}$ items.

3—Grouped data of discrete & continuous variate:

In the case of the grouped data, the arrangement of the items according to their magnitudes is already done. What is required next to be done, is to calculate.

(i) The cumulative frequencies,

(ii) To locate the median-class, which is $\left\{\frac{n+1}{2}\right\}^{\text{th}}$ c.f.

(iii) To determine the value of the median by applying the formula—

$$Md = L + \frac{m-c}{f} \times i$$

where Md —→ median, $m = \frac{n+1}{2}$

L —→ lower limit of the median class

i —→ class-interval of the median class,

f —→ frequency of the median class

c —→ cumulative frequency of the class following the median class

For illustration, consider the following frequency distribution to find out the median—

class	frequency (f)
0—5	4
5—10	6
10—15	10
15—20	16
20—25	12
25—30	8
30—35	4

To calculate the median for the above data, first we prepare the following table for c.f.

class	frequency (f)	c . f .
0—5	4	4
5—10	6	10
10—15	10	20
15—20	16	36
20—25	12	48
25—30	8	56
30—35	4	60

Here, we want the value of $\left(\frac{60+1}{2}\right)^{\text{th}} = 30.5^{\text{th}}$ item, which is located in the class 15—20. Then the median is determined by the formula—

$$\text{Md} = L + \frac{m-c}{f} \times i \quad \text{where } m = \frac{n+1}{2} = 30.5$$

$$= 15 + \frac{30.5 - 20}{16} \times 5$$

$$= 15 + 3.28$$

\therefore median = 18.28 approximately.

$$L = 15$$

$$c = 20$$

$$f = 16$$

$$i = 5$$

Example No. (8)

What do you mean by the Quartiles ? Calculate the median, the upper and the lower quartiles from the data given below—

% of recovery of sugar on cane	No. of factories	% of recovery of sugar on cane	No. of factories
8.0—8.2	2	9.4—9.6	10
8.2—8.4	5	9.6—9.8	7
8.4—8.6	4	9.8—10.0	6
8.6—8.8	11	10.0—10.2	3
8.8—9.0	11	10.2—10.4	1
9.0—9.2	1	10.4—10.6	1
9.2—9.4	13		
Total			85

Solution—

As the median is such a value of the variate which divides the whole distribution in such a way that half of the observations lie below it and the remaining half above it when the data is arranged in the ascending or descending order of magnitudes.

Similarly, the lower quartile denoted by Q_1 is the value of the variate which divides the distribution in such a manner that one quarter of the observations lie below it and the remaining 3 quarters above it when the values have been arranged in the ascending order of their magnitudes.

The upper quartile denoted by Q_3 is the value of the variate which divides the distribution in such a way that 3 quarters of the observations lie below it and 1 quarter above it when the data have been arranged in the ascending order of their magnitudes.

Clearly Q_2 is the median. Q_1 , Q_3 and median, are also called the *partitioning values of the distribution*. Other partitioning values are Per tiles, Deciles and Hectiles.

The mathematical formula for the computation of quartiles for the grouped data are—

$$Q_1 = L + \frac{q_1 - c}{f} \times i$$

where $L \rightarrow$ lower limit of Q_1 class

$$q_1 = \frac{n+1}{4}$$

$c \rightarrow$ c.f. of the class following the Q_1 class

$f \rightarrow$ frequency of the Q_1 class

$i \rightarrow$ class interval of the Q_1 class

and

$$Q_3 = L + \frac{q_3 - c}{f} \times i$$

$$\text{where } q_3 = \frac{3(n+1)}{4}$$

$L \rightarrow$ lower limit of Q_3 class

$f \rightarrow$ frequency of Q_3 class

$i \rightarrow$ class interval of Q_3 class

$c \rightarrow$ c.f. of the class following the Q_3 class

The Q_1 class lies opposite of $\left(\frac{n+1}{4}\right)$ c.f. and Q_3 class lies opposite of $\frac{3(n+1)}{4}$ c.f. For the present example, we have the following table—

% recovery of sugar on cane x	No. of factories (f)	c. f .
8.0 - 8.2	2	2
8.2 - 8.4	5	7
8.4 - 8.6	4	11
8.6 - 8.8	11	22
8.8 - 9.0	11	33
9.0 - 9.2	11	44
9.2 - 9.4	13	57
9.4 - 9.6	10	67
9.6 - 9.8	7	74
9.8 - 10.0	6	80
10.0 - 10.2	3	83
10.2 - 10.4	1	84
10.4 - 10.6	1	85

Measures of Central tendency (averages) and dispersion 107

$$Md = L + \frac{m - c}{f} \times i \quad \text{where } m = \frac{85 + 1}{2} = 43$$

Hence the median class is 9.0—9.2 and so $L = 9.0$,
 $i = 0.2$, $f = 11$, $c = 33$

$$\begin{aligned} \therefore Md &= 9.0 + \frac{43 - 33}{11} \times 0.2 \\ &= 9.0 + 0.1818 \end{aligned}$$

$$\therefore \text{median} = 9.1818$$

$$\text{The lower quartile } Q_1 = L + \frac{q_1 - c}{f} \times i$$

$$\text{where } Q_1 = \frac{85 + 1}{4} = 21.5$$

Hence the lower quartile class is 8.6—8.8 and so $L = 8.6$
 $i = 0.2$, $f = 11$, $c = 11$

$$\begin{aligned} \therefore Q_1 &= 8.6 + \frac{21.5 - 11}{11} \times 0.2 \\ &= 8.6 + 0.1818 \\ Q_1 &= 8.7818 \end{aligned}$$

$$\text{The upper quartile } Q_3 = L + \frac{q_3 - c}{f} \times i$$

$$\text{where } q_3 = \frac{3(85 + 1)}{4} = 64.5$$

Hence the upper quartile class is 9.4—9.6 and so
 $L = 9.4$, $i = 0.2$, $f = 10$, $c = 57$

$$\begin{aligned} \therefore Q_3 &= 9.4 + \frac{64.5 - 57}{10} \times 0.2 \\ &= 9.4 + 0.15 \end{aligned}$$

$$\therefore Q_3 = 9.55$$

ExampleNo. (10)

The classification of 75 cows has been done according to one day milk in the following table. Find the median ?

class interval (Milk in lbs)	No. of cows
8—10	4
10—12	8
12—14	12
14—16	25
16—18	15
18—20	8
20—22	3

Solution:—

We first calculate the cumulative frequencies which are given in the following table—

class interval (x)	frequency (f)	c . f .
8—10	4	4
10—12	8	12
12—14	12	24
14—16	25	49
16—18	15	64
18—20	8	72
20—22	3	75

The median is computed by the formula—

$$Md = L + \frac{m-c}{f} \times i$$

$$\text{where } m = \frac{n+1}{2} = \frac{75+1}{2} = 38$$

Hence the median lies in the class 14—16 and so $L=14$, $f=25$, $c=24$, $i=2$

$$Md = 14 + \frac{38-24}{25} \times 2$$

$$= 14 + 1.12 = 15.12$$

$$\therefore \text{Median} = 15.12$$

Example No. 11

The following table gives the marks obtained by a batch of candidates in a certain examination of History and politics. In which subject is the level of knowledge of the candidates higher ? Give reasons ?

Roll No.	History	Politics	Roll No.	History	Politics
1	42	46	9	40	30
2	24	20	10	62	61
3	38	41	11	55	50
4	35	43	12	54	63
5	30	25	13	52	45
6	45	54	14	47	56
7	58	47	15	43	58
8	50	36			(M.Sc. Agra 1955)

Solution:—

To find out the subject in which the knowledge of students is higher, we calculate the medians of the marks obtained in the two subjects. In the subject, for which the median is higher, the level of knowledge is higher.

For this, we arrange the marks in the ascending order of magnitudes in each subject.

History : 24, 30, 35, 38, 40, 42, 43, 45, 47, 50, 52, 54, 55, 58, 62

Politics : 20, 25, 30, 36, 41, 43, 45, 46, 47, 50, 54, 56, 58, 61, 63

The median will be the size of $\left\{ \frac{n+1}{2} \right\}^{\text{th}}$ item.

=size of 8th item, since $n=15$

Median for History-marks=45

Median for Politics-marks=46

Hence, the level of knowledge of the candidates is higher in Politics than in the history.

Example No. (12)

Define Mode ? Illustrate its method of calculation ?

Solution:—

Mode : The most frequent or the most popular value of a variate is called the mode. It is repeated at the greatest no of times. In popular language, when we speak the average student or average rent, we generally imply the modal student or the modal rent. It is easy to locate, as it lies at the highest frequency. But if there are irregularities in the frequency distribution the position of the mode could become indefinite. In such cases, the process of grouping will be applicable. In a discrete series the size of the variable which has the max. frequency is the mode; while in a continuous series the mode will be located by interpolation in the modal group by the formula

$$Mo = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i$$

where L means the lower limit of the modal class

Δ_1 —→ Stands for the difference between the frequency of the modal class and that of the class which follows the modal class

Δ_2 —→ Stands for the difference between the frequency of the modal class and that of the the class which preceeds the modal class.

i —→ Stands for the class interval of the modal class.

Let us consider the following age-distribution of the candidates appearing at the matriculation examination of Patna University in 1937—

Age-group (years)	No. of stud.	Age-group (years)	No. of stud.
12—13	5	17—18	980
13—14	48	18—19	981
14—15	189	19—20	794
15—16	303	20—21	515
16—17	522	21—22	474
Total			4811

Here the modal class is 18-19 corresponding to the highest frequency 981 and so,

$$L=18, i=19-18=1, \Delta_1=981-980=1, \Delta_2=981-794=187$$

$$\therefore Mo=18+\frac{1}{1+187} \times 1$$

$$=18+0.005=18.005$$

$$\therefore \text{Mode}=18.005 \text{ years.}$$

ExampleNo. (13)

Define dispersion ? What are the different methods of measuring it, describe each briefly ? (U.P. Board, 1962)

Solution:—

Having been known the value of the average of a series, we try to compute some statistics which can give an idea how the values of the variate are scattered around the central value. This variation of the data from the central value is called the *dispersion or Scatter or Spread or variability*.

Thus, the dispersion can be defined as the extent to which the magnitudes of the items differ from the central value.

The following are the measures of dispersion—

- (i) Range
- (ii) Quartile deviation
- (iii) Mean deviation from (a) average (mean)
(b) median
- (iv) Standard deviation

(i) Range : It is defined as the difference between the greatest and the least magnitude of the items. It is very easy to calculate but its use as a measure of dispersion is very rare. Because it does not depend upon all the observations (items) and in its computation no consideration is given to the value of the central tendency.

(ii, Quartile deviation : (Semi Inter quartile range)
It is half of the quartile range i.e.

$$Q.D. = \frac{Q_3 - Q_1}{2}$$

It is a better measure of dispersion than the range and is oftenly used in elementary descriptive statistics. But due to its incapability of algebraic treatments, it is not used in the

Advance-theory. Another reason is that it gives an idea of dispersion only of those items which lie between Q_1 & Q_3 and so most of the items have no effect on the computation of dispersion.

(iii) **Mean deviation** (average deviation) : As the computations of range & quartile deviation do not depend upon all the observation, so these measures of dispersion can not be said as the satisfactory measures of dispersion. One measure free from this objection is the mean deviation which is defined as the arithmetic average of the absolute deviations of the items from any measure (mean or median) of central tendency. Mathematically we write—

$$\text{Mean deviation} = \frac{\sum f|d|}{n},$$

where n is the total number of items,

$f \rightarrow$ Stands for the corresponding frequency of the variate x

and $|d| \rightarrow$ Stands for the absolute deviation of x from its average.

If the deviations are taken from mean, then it is called *Mean Deviation about mean* and if deviations are taken from the median, then it is called *Mean Deviation about median*. The mean deviation also suffers from one drawback that is of incapability of further algebraic treatments.

For the illustration of the procedure, we calculate the mean deviation about mean from the following data—

$x :$	2.5,	7.5,	12.5,	17.5,	22.5,	27.5,	32.5
$f :$	4,	6,	10,	16,	12,	8,	4

To compute the mean deviation about the mean, we form the following table—

(1) x	(2) f	(3) f x	(4) d=x-M	(5) d	(6) f d
2.5	4	10.0	-15.5	15.5	62.0
7.5	6	45.0	-10.5	10.5	63.0
12.5	10	125.0	-5.5	5.5	55.0
17.5	16	280.0	-0.5	0.5	8.0
22.5	12	270.0	+4.5	4.5	54.0
27.5	8	220.0	+9.5	9.5	76.0
32.5	4	130.0	+14.5	14.5	58.0
Γ totals	60=Σf	1080 =Σfx			376.0 =Σf d

$$\text{Mean } \bar{X} \text{ or } M = \frac{\Sigma fx}{n} = \frac{1080}{60} = 18$$

First of all, we calculate the mean $\bar{X} = 18$, then we take the deviations of all values of the variate of colⁿ. (1) from their actual mean 18 as shown in colⁿ. (4), by 'd'. Again in colⁿ. (5), we write the absolute values of 'd' and in (6)th the product of (2) & (5) with respective figures.

$$\text{Mean deviation (about mean)} = \frac{\Sigma f |d|}{n} = \frac{376}{60} = 6.266$$

If we want to calculate the mean deviation about the median then first we find the median for the data and then take the deviations of the values from this median. Similarly, the mean deviation about any average can be found

by first computing that desired average for the given data and then obtaining the deviations of the values from that average. The rest of the procedure is the same as indicated above.

(iv) Standard deviation : Among all the measures of dispersion, the S.D. is most widely used because of the facts that

- (1) its computation is based on all the observations,
- (2) the deviations are taken from the mean and so the sum of squares of the deviations is minimum,
- (3) it is capable of algebraic treatments.

The S.D. is the square root of the mean of the squares of deviations from the arithmetic mean. Symbolically, the Greek letter sigma (σ) denotes the standard deviation.

$$\sigma = \sqrt{\frac{\sum (x - \bar{X})^2}{n}} = \sqrt{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2}$$

and $\sigma = \sqrt{\frac{\sum f(X - \bar{X})^2}{n}} = \sqrt{\frac{\sum fx^2}{n} - \left(\frac{\sum fx}{n}\right)^2}$

We can take the deviations from an assumed mean also in place of the actual mean for our convenience. Then the formula for S. D. will be

$$\sigma = \sqrt{\frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2}$$

and $\sigma = \sqrt{\frac{\sum fd^2}{n} - \left(\frac{\sum fd}{n}\right)^2}$ for a frequency distribution.

For illustration, we summarize the steps in the computation of S.D. by taking the following simple series of values—

X : 192, 288, 236, 229, 184, 230, 348, 291, 330, 243

- (1) Choose the assumed mean 'A',
- (2) Take the deviations 'd' from this assumed mean,

3. Square the deviations *i. e.* calculate ' d^2 '.

The results of the above steps for the given series are tabulated

\times	$d = X - A$ $A = 260$	d^2
192	-68	4624
288	+28	784
236	-24	576
229	-31	961
184	-76	5776
260	0	0
348	+88	7744
291	+31	961
330	+70	4900
243	-17	289
Totals	$+1 = \Sigma d$	$26615 = \Sigma d^2$

The S. D. is given by

$$\begin{aligned}
 \sigma &= \sqrt{\left\{ \frac{\Sigma d^2}{n} - \left(\frac{\Sigma d}{n} \right)^2 \right\}} \\
 &= \sqrt{\left\{ \frac{26615}{10} - \left(\frac{1}{10} \right)^2 \right\}} \\
 &= \sqrt{(2661.5 - .01)} \\
 &= \sqrt{(2661.49)} \\
 \sigma &\approx 51.59.
 \end{aligned}$$

To complete the S. D. in the case of frequency distribution, the following steps are made—

- (1) Choose the assumed mean, ' A '.
- (2) take the deviations $d = (X - A)$ from this assumed mean.
- (3) calculate the squares of d .
- (4) compute the products (fd) in a column.
- (5) compute the products (fd^2) in another column.

For illustration of the procedure, we calculate the S. D. from the following data—

Height in cms. (X) : 60, 61, 62, 63, 64, 65, 66, 67, 68.

No. of plants (f) ; 2, 0, 15, 29, 25, 12, 10, 4, 3.

The computational work is done in the following tabular form—

$$A=64$$

X	f	d=X-A	d ²	fd	fd ²
60	2	-4	16	-8	32
61	0	-3	9	0	0
62	15	-2	4	-30	60
63	29	-1	1	-29	29
64	25	0	0	0	0
65	12	1	1	+12	12
66	10	2	4	+20	40
67	4	3	9	+12	37
68	3	4	16	+12	48
Totals	100			-11 = Σfd	257 = Σfd^2

The S. D. is given by—

$$\sigma = \sqrt{\left\{ \frac{\Sigma fd^2}{n} - \left(\frac{\Sigma fd}{n} \right)^2 \right\}} = \sqrt{\left\{ \frac{257}{100} - \left(\frac{-11}{100} \right)^2 \right\}} = \sqrt{(2.57 - .0121)} \\ = \sqrt{(2.5579)} \\ \approx 1.59$$

Computation of the S. D. in the case of frequency distribution of the grouped data with unequal Interval.

In this case, the classes are replaced by their mid-values and the rest procedure remains the same as explained above.

In the case of a grouped data when the class-intervals are equal, a more convenient formula for calculating the S. D. is.

$$\sigma = i \times \sqrt{\left\{ \frac{\Sigma f \xi^2}{n} - \left(\frac{\Sigma f \xi}{n} \right)^2 \right\}}, \text{ where } i \text{ is the class-interval and} \\ \xi = \frac{d}{i}$$

This is known as the *method of step-deviation*.

The steps according to this method are given below—

- (1) choose the assumed mean 'A',
- (2) take the deviations $d=(X-A)$ from this assumed mean,
- (3) divide the deviations by the class interval say i and call the value $\frac{d}{i} = \xi$

- (4) calculate the squares of ξ
- (5) compute the products $f \xi$, and
- (6) finally compute the products $f \xi^2$.

For illustration, consider the following example—

Age group (X) : 20—30, 30—40, 40—50, 50—60, 60—70,
(Years) 70—80, 80—90.

No. of persons (f) : 3, 61, 132, 153, 140, 51, 2

The computational work for S. D. is shown in the following tabular form—

$$A=55, \quad i=10$$

Age group (Years)	Mid value X	f	d= (X-A)	$\xi=d/i$	f ξ	f ξ^2
20—30	25	3	-30	-3	-9	27
30—40	35	61	20	+2	122	244
40—50	45	132	-10	-1	-132	132
50—60	55	153	0	0	0	0
60—70	65	140	+10	+1	+140	140
70—80	75	51	+20	+2	+102	204
80—90	85	2	+30	+3	+6	18
Totals		542= Σf			-15= $\Sigma f \xi$	765= $\Sigma f \xi^2$

The S. D. is given by

$$\begin{aligned} \sigma &= i \times \sqrt{\left\{ \frac{\Sigma f \xi^2}{n} - \left(\frac{\Sigma f \xi}{n} \right)^2 \right\}} \\ &= 10 \times \sqrt{\left\{ \frac{765}{542} - \left(\frac{-15}{542} \right)^2 \right\}} \end{aligned}$$

$$\begin{aligned}
 &= 10 \times \sqrt{(1.4114 - .0008)} = 10 \times \sqrt{(1.4106)} \\
 &= 10 \times 1.187 \\
 &\approx 11.87 \\
 \sigma &= 11.87 \text{ approximately.}
 \end{aligned}$$

Exp. 14. Prove that $\sum x^2 = \sum X^2 - \frac{(\sum X)^2}{n}$

Where x is the deviation of an observation X from the arithmetic mean \bar{X} of the sample and n is the total number of observations in the sample.

(M. Sc. Ag. Agra, 1963)

Sol. We have $x = X - \bar{X}$, where $\bar{X} = \frac{\sum X}{n}$

$$\begin{aligned}
 x^2 &= (X - \bar{X})^2 \\
 \text{and } \sum x^2 &= \sum (X - \bar{X})^2 \\
 &= \sum [X^2 + \bar{X}^2 - 2X\bar{X}] \\
 &= \sum X^2 + \sum \bar{X}^2 - 2\bar{X}\sum X \\
 &= \sum X^2 + n\bar{X}^2 - 2n\bar{X}^2 \quad \because \sum X = n\bar{X} \\
 &= \sum X^2 - n\bar{X}^2 \\
 &= \sum X^2 - n \cdot \left(\frac{\sum X}{n} \right)^2 \\
 \sum x^2 &= \sum X^2 - \frac{(\sum X)^2}{n}
 \end{aligned}$$

Hence Proved.

Exp. 15. Compute the S. D. of the following data by means of the formulac—

$$(a) \quad \sigma^2 = \frac{\sum f(X - \bar{X})^2}{n}$$

$$(b) \quad \sigma^2 = \frac{\sum fX^2}{n} - \left(\frac{\sum fX}{n} \right)^2$$

$$\text{and (c) } \sigma^2 = \frac{\sum fd^2}{n} - \left(\frac{\sum fd}{n} \right)^2$$

$x :$	12,	13,	14,	15,	16,	17,	18,	20
$f :$	4,	11,	32,	21,	15,	8,	5,	4

Sol. : (a) For the computation of the S. D., the table is given below—

$$\bar{X} = 15$$

X	f	fX	d = (X - \bar{X})	(X - \bar{X}) ²	f(X - \bar{X}) ²
12	4	48	-3	9	36
13	11	143	-2	4	44
14	32	448	-1	1	32
15	21	315	0	0	0
16	15	240	1	1	15
17	8	136	2	4	32
18	5	90	3	9	45
20	4	80	5	25	100
Totals	$\Sigma f = 100$	$\Sigma fX = 1500$			$\Sigma f(X - \bar{X})^2 = 304$

$$\bar{X} = \frac{\Sigma fX}{\Sigma f} = \frac{1500}{100} = 15$$

Thus the S. D. is given by

$$\sigma = \sqrt{\left\{ \frac{\Sigma f(X - \bar{X})^2}{n} \right\}} = \sqrt{\frac{304}{100}} = \sqrt{3.04}$$

∴ $\sigma = 1.74$ approximately.

(b) To compute the S. D., the following table is framed—

X	f	X ²	fX	fX ²
12	4	144	48	576
13	11	169	143	1859
14	32	196	448	6272
15	21	225	315	4725
16	15	256	240	3840
17	8	289	136	2312
18	5	324	90	1620
20	4	400	80	1600
Totals	$100 = \Sigma f$		$\Sigma fX = 1500$	$22,804 = \Sigma fX^2$

The S. D. is given by

$$\sigma = \sqrt{\left\{ \frac{\sum fX^2}{n} - \left(\frac{\sum fX}{n} \right)^2 \right\}} = \sqrt{\left\{ \frac{22,804}{100} - \left(\frac{1500}{100} \right)^2 \right\}}$$

$$= \sqrt{(228.04 - 225)}$$

$$\therefore \sigma = \sqrt{3.04} = 1.76 \text{ approximately.}$$

(c) The S. D. is computed with the help of the following table-

$$A = 16$$

X	f	d=(X-A)	d ²	fd	fd ²
12	4	-4	16	-16	64
13	11	-3	9	-33	99
14	32	-2	4	-64	128
15	21	-1	1	-21	21
16	15	0	0	0	0
17	8	1	1	8	8
18	5	2	4	10	20
20	4	4	16	16	64
Totals	100 = $\sum f$			-100 = $\sum fd$	404 = $\sum fd^2$

$$\sigma = \sqrt{\left\{ \frac{\sum fd^2}{n} - \left(\frac{\sum fd}{n} \right)^2 \right\}} = \sqrt{\left\{ \frac{404}{100} - \left(\frac{-100}{100} \right)^2 \right\}}$$

$$= \sqrt{(4.04 - 1)}$$

$$= \sqrt{3.04}$$

$$\sigma = 1.76 \text{ approximately.}$$

The calculations have been considerably reduced by using

$$= \sqrt{\left\{ \frac{\sum fd^2}{n} - \left(\frac{\sum fd}{n} \right)^2 \right\}} \text{ and so it is called short cut formula.}$$

Exp. 16 Calculate the median and the quartiles using

(a) mathematical method, and

(b) graphical method, of the following data—

Class	Frequency
5—7	4
7—9	6
9—11	10
11—13	12
13—15	6
15—17	5
17—19	4

Sol. : First we prepare the cumulative frequency table—

Class	frequency	c. f.
5—7	4	4
7—9	6	10
9—11	10	20
11—13	12	32
13—15	6	38
15—17	5	43
17—19	4	47 = n

We have

$$md = L + \frac{m - c}{f} \times i$$

$$\text{where } m = \frac{n+1}{2} = 24$$

So, median lies in the class 11—13

$$Md = 11 + \frac{24 - 20}{12} \times 2$$

$$= 11 + 0.66$$

$$\therefore \text{Median} = 11.66$$

The lower quartile is,

$$Q_1 = L + \frac{q_1 - c}{f} \times i \text{ where } q_1 = \frac{n+1}{4} = \frac{47+1}{4} = 12 \text{ and so,}$$

$$= 9 + \frac{12 - 10}{10} \times 2$$

$$\begin{aligned}
 &= 9 + 0.4 \\
 &= 9.4 \\
 Q_1 &= 9.4
 \end{aligned}$$

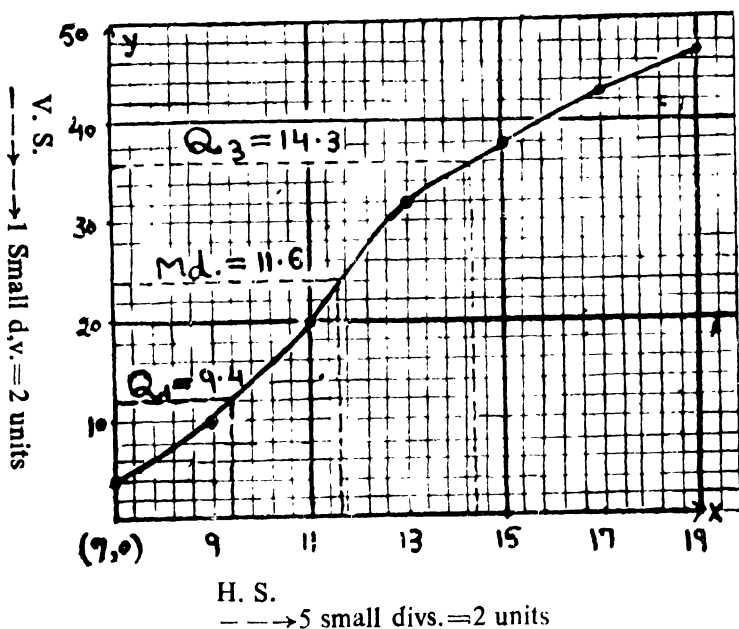
The upper quartile is, $Q_3 = h + \frac{q_3 - c}{f} \times i$

$$\text{where } q_3 = \frac{3(n+1)}{4} = 3 \times 12 = 36 \text{ and}$$

$$\text{so } Q_3 = 13 + \frac{36 - 32}{6} \times 2 = 13 + 1.33 = 14.33 \quad \therefore Q_3 = 14.33$$

(b) To determine any partition-value *i. e.* median and quartiles etc., we first draw a *c. f.* curve and then draw perpendicular

C. F. Curve (Ogive)



Graph No. (1)

lines to the *y*-axis corresponding to the partition item *i. e.* m , q_1 & q_3 etc. as the case may be. Next, from the points where these lines cut the ogive, draw the lines perpendicular to the *x*-axis. The distances between the origin and the foot of the perpendiculars on

the x -axis are the partition-values. In the present example, we take points at 12, 24, 36 on y -axis and from these points the lines parallel to the x -axis are drawn. The points are noted where these lines cut the ogive. Finally the perpendiculars are drawn from these points on the x -axis. The abscissae of these points give the desired partition-values.

	Mathematical value	Graphical value
Results : $Q_1=9.4$,	9.4
$Md=11.66$,	11.6
$Q_3=14.33$,	14.3

Exp. 17. (a) If $y=ax+b$ and variance of x is σ_x^2 , then show that

$$\sigma_y^2 = a^2 \cdot \sigma_x^2.$$

(b) If x and y are two independent variates with their respective variances σ_x^2 and σ_y^2 and a new variate z is defined by the relation $z=ax+by$, then show that $\sigma_z^2 = a^2\sigma_x^2 + b^2\sigma_y^2$.

(c) If the standard deviations of the two variates x and y be 0.5 and 1.2 respectively, find the S. Ds. of the variates

$$(x-y), (x+y) \text{ and } (3x-4y) ?$$

Sol. (a) We know that

$$\begin{aligned} \sigma_y^2 &= \frac{\sum y^2}{n} - \left(\frac{\sum y}{n} \right)^2 \\ &= \frac{\sum (ax+b)^2}{n} - \left[\frac{\sum (ax+b)}{n} \right]^2 \\ &= \frac{\sum (a^2x^2 + 2abx + b^2)}{n} - \left[\frac{\sum ax}{n} + b \right]^2 \\ &= a^2 \frac{\sum x^2}{n} + 2ab \frac{\sum x}{n} + b^2 - \left[\left(\frac{\sum ax}{n} \right)^2 + b^2 + 2ab \frac{\sum x}{n} \right] \\ &= a^2 \frac{\sum x^2}{n} - a^2 \left(\frac{\sum x}{n} \right)^2 \\ &= a^2 \left\{ \frac{\sum x^2}{n} - \left(\frac{\sum x}{n} \right)^2 \right\} \\ \therefore \sigma_y^2 &= a^2 \cdot \sigma_x^2. \end{aligned}$$

Hence proved.

(b) We have

$$\begin{aligned} \sigma_z^2 &= \frac{\sum z^2}{n} - \left(\frac{\sum z}{n} \right)^2 \\ &= \frac{\sum (ax+by)^2}{n} - \left[\frac{\sum (ax+by)}{n} \right]^2 \\ &= \frac{\sum (a^2x^2 + b^2y^2 + 2abxy)}{n} - \left\{ \frac{a\sum x}{n} + \frac{b\sum y}{n} \right\}^2 \end{aligned}$$

$$= a^2 \frac{\sum x^2}{n} + 2ab \frac{\sum xy}{n} + b^2 \frac{\sum y^2}{n} - \left\{ a^2 \left(\frac{\sum x}{n} \right)^2 + b^2 \left(\frac{\sum y}{n} \right)^2 + 2ab \frac{\sum x}{n} \frac{\sum y}{n} \right\}$$

or

$$\sigma_z^2 = a^2 \left\{ \frac{\sum x^2}{n} - \left(\frac{\sum x}{n} \right)^2 \right\} + b^2 \left\{ \frac{\sum y^2}{n} - \left(\frac{\sum y}{n} \right)^2 \right\} + 2ab \left\{ \frac{\sum xy}{n} - \bar{x} \cdot \bar{y} \right\}$$

$$= a^2 \cdot \sigma_x^2 + b^2 \sigma_y^2 + \frac{2ab}{n} \{ \sum xy - n\bar{x}\bar{y} \}$$

$$= a^2 \sigma_x^2 + b^2 \sigma_y^2 + \frac{2ab}{n} \sum (x - \bar{x})(y - \bar{y})$$

$$\therefore \sigma_z^2 = a^2 \sigma_x^2 + b^2 \sigma_y^2 \quad ()$$

Because $\sum (x - \bar{x})(y - \bar{y}) = 0$ as x and y are independent variates.

$$\therefore \sigma_z^2 = a^2 \sigma_x^2 + b^2 \sigma_y^2.$$

Hence proved

$$(c) V(x+y) = V(x) + V(y)$$

$$= 0.5 + 1.2$$

$$= 1.7.$$

when x and y are indep.and $a = 1, b = 1$

$$\therefore S. D. (x+y) = \sqrt{1.7}$$

or

$$\sigma_{(x+y)} = 1.3 \text{ approximately.}$$

Similarly, $V(x-y) = V(x) + V(y)$, where $a = 1, b = -1$.

$$= 0.5 + 1.2$$

$$= 1.7$$

$$\therefore S. D. (x-y) = \sqrt{1.7}$$

$$\sigma_{(x-y)} = 1.3 \text{ approximately.}$$

Also, $V(3x-4y) = 9 \times V(x) + 16 \times V(y)$

$$= 9 \times 0.5 + 16 \times 1.2$$

$$= 4.5 + 19.2$$

$$= 23.7$$

as $a = 3$ $b = -4$

$$S. D. (3x-4y) = \sqrt{23.7}$$

or

$$\sigma_{(3x-4y)} = 4.86 \text{ approximately.}$$

Exp. 18. Write notes on :

- (a) Variance,
- (b) Standard error,
- (c) Coefficient of variation

Sol. (a) Variance : It is the square of the standard deviation i.e.

$$\text{Variance } = (\text{S. D.})^2 = \frac{\sum (X - \bar{X})^2}{n}$$

where $X \rightarrow$ represents the variate values

$\bar{X} \rightarrow$ the mean of the variate values

$n \rightarrow$ the total number of items.

Thus, variance is the arithmetic mean of the squares of the deviations from the mean. It is denoted by σ^2 and is widely used in the statistical analysis of field experiments.

(b) Standard error : It is the standard deviation of any statistic calculated on the basis of sample observations. It is widely used in the testing of statistical hypotheses. The S. E. of the mean is given by $\frac{\text{S. D.}}{\sqrt{n}}$, where n is the total number of observations, on which the mean has been calculated i.e.

$$\text{S. E.} = \frac{\sigma}{\sqrt{n}}$$

(c) Coefficient of variation : The S. D. and other measures of dispersion are the absolute measures of dispersion and are expressed in the same units in which the observations are given and hence cannot be used to compare the variations in the two given series which differ in their units and averages. To compare the variations of two such series, we require some relative measure of dispersion. *The coefficient of variation is such a measure which is defined as the ratio of the S. D. to the mean expressed in percentage.* Symbolically,

$$\text{C. V.} = \frac{\sigma}{M} \times 100.$$

This statistic was developed by **Pearson**. In order to compare the variations in the two series, we compute the coefficients of variation for each of them. The series for which the C. V. is higher, will vary more than the other.

Exp. 19. The scores of two golfers for 10 rounds each are :

A : 58, 59, 60, 54, 65, 66, 52, 75, 69, 52

B : 84, 56, 92, 65, 86, 78, 44, 54, 73, 68

which may be regarded as the more consistent player ?

Sol. In the present problem, we have to compute the coefficients of variation in the two cases and then compare them. The one, which has less variability than the other, will be called as the more consistent player ?

The computation is made in the following tabular form :

Rounds	Golfer A			Golfer B		
	Scores (X)	$(X - \bar{X})$	$(X - \bar{X})^2$	Scores (Y)	$(Y - \bar{Y})$	$(Y - \bar{Y})^2$
1	58	-3	9	84	+14	196
2	59	-2	4	56	-14	196
3	60	-1	1	92	+22	484
4	54	-7	49	65	-5	25
5	65	+4	16	86	+16	256
6	66	+5	25	78	+8	64
7	52	-9	81	44	-26	676
8	75	+14	196	54	-16	256
9	69	+8	64	73	+3	9
10	52	-9	81	68	-2	4
Totals	610 = ΣX	0	526 = $\Sigma (X - \bar{X})^2$	700 = ΣY	0	$\Sigma (Y - \bar{Y})^2 = 2166$

$$\bar{X} = \frac{\Sigma X}{n} = \frac{610}{10} = 61$$

$$\bar{Y} = \frac{\Sigma Y}{n} = \frac{700}{10} = 70$$

$$\begin{aligned} \text{S. D. } (\sigma_x) &= \sqrt{\left(\frac{\Sigma (X - \bar{X})^2}{n} \right)} \\ &= \sqrt{\left(\frac{526}{10} \right)} \end{aligned}$$

$$\begin{aligned} \text{S. D. } (\sigma_y) &= \sqrt{\left(\frac{\Sigma (Y - \bar{Y})^2}{n} \right)} \\ &= \sqrt{\left(\frac{2166}{10} \right)} \end{aligned}$$

$$= \sqrt{(52.6)}$$

$$= 7.252 \text{ approx.}$$

$$\text{C. V.}_A = \frac{\sigma}{M} \times 100 = \frac{7.252}{61} \times 100$$

$$= 11.88$$

$$= \sqrt{(216.6)}$$

$$= 14.717 \text{ approx.}$$

$$\text{C. V.}_B = \frac{14.717}{70} \times 100$$

$$= 21.024$$

Sol. Comparing the two C. Vs., we conclude that the player A is more constant than player B.

Exp. 20. (a) Explain the practical utility of the coefficient of variation in the Biological research ?

(b) The average heights of ten years old and 18 years old girls were reported to be 74.4 and 161.0 cms. respectively with standard deviations 2.64 and 6.12 cms. respectively. Which data varies more ?
(M Sc. Ag. Agra 1963)

Sol. (a) It has been answered in the above question.

(b) C. V. for the heights of 10 years old girls is

$$\text{C. V.}_1 = \frac{2.64}{74.4} \times 100$$

$$= 3.548.$$

C. V. for the heights of 18 years old girls is

$$\text{C. V.}_2 = \frac{6.12}{161.0} \times 100$$

$$= 3.801.$$

Comparing the two C. Vs., we conclude that the heights of the 18 years old girls show the greater variability than that of the 10 years old girls.

Exp. 21. A random sample from a biological population is given below :

Observation No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
No. of grains/ear head	19	30	11	31	32	36	39	24	12	57	39	29	34	53	33

In the above series, sum of all the observations is 479 and sum of squares of deviations from the mean is 2272.93.

(a) Calculate the following of the above sample.

Mean, variance, standard deviation and standard error.

(b) What is an array ? Illustrate your answer with the series given above.

(c) In samples of 15-20 items, the range is on the average about 3.5 times the standard deviation. Verify this statement from the sample given above? (*M. Sc. Ag. Agra, 1964*)

Sol. (a) We have

$$\begin{aligned}n &= 15 \\ \Sigma X &= 479 \\ \Sigma (X - \bar{X})^2 &= 2272.93.\end{aligned}$$

$$\text{Hence, the mean } (\bar{X}) = \frac{\Sigma X}{n} = \frac{479}{15} = 31.93 \text{ grains/earhead}$$

$$\text{Variance } (\sigma_x^2) = \frac{\Sigma (X - \bar{X})^2}{n} = \frac{2272.93}{15} = 151.5286$$

$$\begin{aligned}\text{S. D. } (\sigma_x) &= \sqrt{\left(\frac{\Sigma (X - \bar{X})^2}{n}\right)} = \sqrt{\left(\frac{2272.93}{15}\right)} \\ &= \sqrt{(151.5286)} = 12.309\end{aligned}$$

$$\begin{aligned}\therefore \text{E. (of mean)} &= \frac{\text{S. D.}}{\sqrt{n}} = \frac{\sqrt{\left(\frac{\text{variance}}{n}\right)}}{\sqrt{n}} = \sqrt{\left(\frac{151.5286}{15}\right)} \\ &= \sqrt{(10.1019)} \\ &= 3.17 \text{ approx.}\end{aligned}$$

(b) **Array :** An orderly arrangement of the variate values is called an array. If the variate values i.e. the number of grains/earhead are arranged either in ascending or descending order of their magnitudes, the resulting series will be called an array. The practical utility of an array lies in the calculation of partition values.

(c) In the given sample, the lowest number of grains per earhead is 11 and highest is 57, so the range will be given by :

$$R = 57 - 11 = 46 \text{ grains/earhead}$$

We have the S. D. = 12.309

$$\begin{aligned}\text{so } 3.5 \times \text{S. D.} &= 3.5 \times 12.309 \\ &= 43.08\end{aligned}$$

which is nearly equal to the range value i.e. 46 grains/earhead.

Thus, we can say that on the average the range is 3.5 times that of the S. D. in the sample given above.

Exp. 22. A random sample from a field experiment is given below :

Obs. No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Hts. of Plants in cms.	23	17	20	19	18	22	16	25	13	15	19	21	23	20	21	22	16

In this sample, the sum of all the observations is 340 and the sum of squares of deviations from the mean is 254.

(a) Calculate six statistical constants, three each of measures of type and measures of variability, from the above sample.

(b) Explain what is an array with the help of the above sample. (H. B. S. Statistical Methods, 1965)

Sol. (a)

$$\begin{aligned}
 n &= 17 \\
 \Sigma X &= 340 \\
 \Sigma (X - \bar{X})^2 &= 254
 \end{aligned}$$

The three measures of central tendency are: Mean, Median, and Mode. In this case, the Mode cannot be found in the present example without the frequencies).

$$\text{Then Mean} (\bar{X}) = \frac{\Sigma X}{n} = \frac{340}{17} = 20.0 \text{ cms.}$$

The median can be found out by arranging the values (heights of plants in cms.) in ascending order of their magnitudes. The arrangement will be as follows :

13, 15, 16, 16, 17, 18, 19, 20, 20, 21, 21, 22, 22, 23, 21, 25, 29

$$\begin{aligned}
 \text{Md} &= \text{size of the } \left(\frac{n+1}{2} \right)^{\text{th}} \text{ item} = \text{size of } \left(\frac{17+1}{2} \right)^{\text{th}} \text{ item} \\
 &= \text{size of 9th item.}
 \end{aligned}$$

The size of the 9th item in the orderly arranged series is 20 and so the

$$\text{Median} = 20.0 \text{ cms.}$$

For the harmonic mean, we know that

$$H.M. = \frac{n}{\sum \frac{1}{X}}$$

and so it is computed with the help of the following table :

S.N.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	Totals
X	23	17	20	29	18	22	16	25	13	15	19	21	23	20	21	22	16	340
$\frac{1}{X}$	·0434	·0588	·0500	·0344	·0555	·0454	·0625	·0400	·0770	·0666	·0526	·0450	·0434	·0500	·0450	·0454	·0625	0·8575

$$\therefore H.M. = \frac{17}{0·8575} = 19·9 \text{ approximately.}$$

Measures of Central Tendency (averages) and Dispersion , 133

The three averages (measures of central tendency) are found to be equal.

For the three measures of dispersion, we take the range, the S. D. and the mean deviation (about the mean) as computed below :

Range (R) = highest value — lowest value

$$= 29 - 13 = 16 \text{ cms.}$$

$$\text{S. D. } (\sigma_a) = \sqrt{\left(\frac{\sum(X - \bar{X})^2}{n}\right)} = \sqrt{\left(\frac{254}{17}\right)} = \sqrt{(14.942353)}$$

$$= 3.865 \text{ cms. approximately.}$$

$$\text{Mean Deviation (M. D.)} = \frac{\sum |d|}{n},$$

where $|d|$ means the absolute value of the deviation from the actual mean \bar{X} i.e. $|d| = |X - \bar{X}|$.

To compute it, we have to frame the following table :

We have $\bar{X}=20.0$ cms.

S. N.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	Totals
x	23	17	20	29	18	22	16	25	13	15	19	21	23	20	21	22	16	340
$d=X-\bar{X}$	3	-3	0	9	-2	2	-4	5	-7	-5	-1	1	3	0	1	2	-4	0
$ d $	3	3	0	9	2	2	4	5	7	5	1	1	3	0	1	2	4	52

$$\therefore \text{Mean deviation} = \frac{52}{17} = 3.0588 \text{ cms. approx.}$$

(about mean)

Hence we have, the three calculated—

Measures of average

Mean = 20.0 cms.

Median = 20.0 cms,

H. M. = 19.9 cms.

Measures of dispersion

Range = 16 cms.

S. D. = 3.865 cms.

M. D. = 3.0588 cms.

(b) The answer has been given in the previous question.

Exp. 23. Account of the number of grains on each of forty earheads of wheat gave the following results—

32	23	25	19	26	17	31	23
29	39	34	26	27	38	27	34
33	27	29	27	31	38	37	39
18	34	35	40	34	43	24	31
28	43	20	33	28	34	29	28

Present the data in the form of a frequency distribution. Calculate the mean and standard deviation of the distribution from the grouped data and state the standard error of mean number of grains per earhead, (M. Sc. Ag. Agra 1958)

Sol. The lowest magnitude is 17 and highest is 43. Taking the lower limit of the 1st class 16, and keeping the class-interval equal to 4, we shall obtain the following seven classes with their respective frequencies. The border lines items are placed in the classes where they are as the lower limits.

Class	Tally marks	Frequency
16—20		3
20—24		4
24—28	≡	7
28—32	≡	8
32—36	≡ ≡	10
36—40	≡	5
40—44		3

To compute the mean, S. D. and the standard error of mean number of grains per earhead, we form the following table—

$A=30$

Mid value X	$d=X-A$	d^2	f	fd	fd^2
18	-12	144	3	-36	432
22	-8	64	4	-32	256
26	-4	16	7	-28	112
30	0	0	8	0	0
34	4	16	10	40	160
38	8	64	5	40	320
42	12	144	3	36	432
Totals	—	—	$40=\Sigma f$	$+20=\Sigma fd$	$1712=\Sigma fd^2$

Mean is given by

$$\bar{X}=A+\frac{\Sigma fd}{n}$$

$$=30+\frac{20}{40}$$

$$=30+0.5$$

$$\bar{X}=30.5 \text{ grains}$$

$$\text{S. D. } (\sigma)=\sqrt{\left\{\frac{\Sigma fd^2}{n}-\left(\frac{\Sigma fd}{n}\right)^2\right\}}=\sqrt{\left\{\frac{1712}{40}-\left(\frac{20}{40}\right)^2\right\}}$$

$$=\sqrt{(42.80-0.25)}=\sqrt{(42.55)}.$$

\therefore S. D. = 6.52 approximately.

S. E. of mean no. of grains/earhead is given by

$$\text{S. E.}=\frac{\text{S. D.}}{\sqrt{n}}=\frac{6.52}{\sqrt{40}}=\frac{6.52}{6.32}=1.03.$$

\therefore Standard error = 1.03 grains. }

Mean = 30.5 grains }

S. D. = 6.52 grains }

S. E. of mean no. of grains/earhead }

= 1.03 grains }

...

...

Ans.

Exp. 24. The length of 50 earheads of wheat are given below in centimetres—

10.4	10.8	8.5	10.9	10.5	9.4	11.1	7.0	8.8	10.6
10.9	11.2	10.5	9.7	10.8	8.4	9.3	11.2	7.0	11.3
12.2	8.9	9.6	11.0	10.4	9.9	9.0	10.5	9.6	11.0
11.5	9.7	10.6	10.4	9.5	8.4	8.7	9.1	11.5	10.6
11.8	7.8	10.1	10.1	8.4	11.0	9.5	9.4	9.8	11.7

(a) Present the data in the form of a frequency distribution by choosing a suitable class-interval.

(b) Calculate the mean and median of the distribution and represent it graphically.
(M. Sc. Agra, 1960)

Sol. (a) The lowest magnitude is 7.0 and highest is 12.2. Taking the lower limit of the first class 7.0 and keeping the class-interval equal to 0.8, we obtain the following seven classes with their respective frequencies written against them. *The border line items are placed in the classes where they are as the lower limits—*

Class	Tally marks	Frequency
7.0—7.8		2
7.8—8.6		5
8.6—9.4	I	6
9.4—10.2		13
10.2—11.0		13
11.0—11.8		9
11.8—12.6		2
Totals	50	50

(b) To calculate the mean of the distribution, we form the following table—

Mid Value X	$d = X - A$	f	fd
7.4	-2.4	2	-4.8
8.2	-1.6	5	-8.0
9.0	-0.8	6	-4.8
9.8	0	13	0
10.6	0.8	13	10.4
11.4	1.6	9	14.4
12.2	2.4	2	4.8
Totals	—	$\Sigma f = 50$	$12.0 = \Sigma fd$

Hence the mean is

$$\bar{X} = A + \frac{\Sigma fd}{n}$$

$$= 9.8 + \frac{12.0}{50}$$

$$= 9.8 + 0.24$$

$$\therefore \bar{X} = 10.04 \text{ centimetres}$$

For the computation of median of the distribution we prepare the following table—

Class	freq. (f)	c. f
7.0—7.8	2	2
7.8—8.6	5	7
8.6—9.4	6	13
9.4—10.2	13	26
10.2—11.0	13	39
11.0—11.8	9	48
11.8—12.6	2	50

Thus the median is

$$Md = L + \frac{m - c}{f} \times i$$

$$\text{where } m = \frac{n+1}{2} = \frac{50+1}{2} = 25.5$$

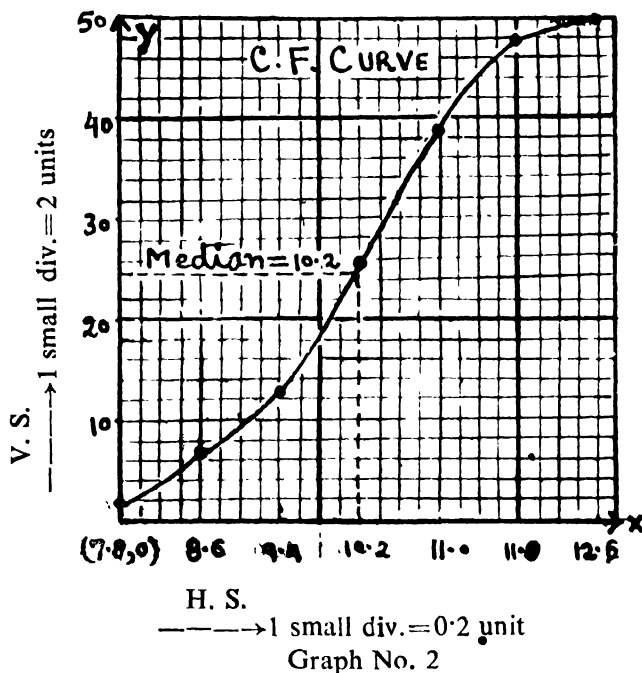
Hence the median lies in the class 9.4—10.2 and so

$$L = 9.4, c = 13, f = 13, i = 0.8$$

$$\begin{aligned}\therefore Md &= 9.4 + \frac{25.5 - 13}{13} \times 0.8 \\ &= 9.4 + 0.7692 \\ &= 10.1692\end{aligned}$$

$$\therefore Md \approx 10.17 \text{ centimetres.}$$

Graphically the median = 10.2 cms. and it is shown in the adjacent graph.



Exp. 25. (a) In a series of 75 items, the total of all the items was found to be 114.55 and their sum of squares 175.7125. Calculate the mean and the standard deviation ?

(b) For the marks obtained by 75 students in a certain test, the sum of the deviation from the assumed mean 17.5 was 330 and the sum of squares of the deviations was 6250. Calculate the mean and the standard deviation ?

Sol. (a) We are given $n = 75$

$$\Sigma x = 114.55$$

$$\Sigma x^2 = 175.7125.$$

$$\therefore \text{Mean} = \frac{\Sigma x}{n} = \frac{114.55}{75} = 1.527$$

$$\begin{aligned}
 \text{S. D. } (\sigma) &= \sqrt{\left\{ \frac{\sum x^2}{n} - \left(\frac{\sum x}{n} \right)^2 \right\}} \\
 &= \sqrt{\left\{ \frac{175 \cdot 7125}{75} - (1 \cdot 527)^2 \right\}} \\
 &= \sqrt{(2 \cdot 3429 - 2 \cdot 3327)} \\
 &= \sqrt{(0 \cdot 0102)}.
 \end{aligned}$$

$\therefore \sigma = 0 \cdot 101$ approximately.

(b) We have,

$$n = 75$$

$$A = 17 \cdot 5$$

$$\sum d = 330$$

$$\sum d^2 = 6250$$

$$\begin{aligned}
 \therefore \text{Mean} &= A + \frac{\sum d}{n} \\
 &= 17 \cdot 5 + \frac{330}{75}
 \end{aligned}$$

$$\therefore \text{Mean} = 17 \cdot 5 + 4 \cdot 4 = 21 \cdot 9.$$

$$\begin{aligned}
 \text{S. D. } (\sigma) &= \sqrt{\left\{ \frac{\sum d^2}{n} - \left(\frac{\sum d}{n} \right)^2 \right\}} \\
 &= \sqrt{\left\{ \frac{6250}{75} - (4 \cdot 4)^2 \right\}} \\
 &= \sqrt{(83 \cdot 3333 - 19 \cdot 36)} = \sqrt{(63 \cdot 9733)} \\
 \therefore \sigma &= 8 \cdot 0 \text{ approximately.}
 \end{aligned}$$

Exp. 26. How many types of frequency functions do you know ; sketch their graphs also ?

Sol. Following are the types of simple frequency distributions :

(i) *Symmetrical Unimodal curves* : This type of the curve is symmetrical about its maximum ordinate. The class-frequencies

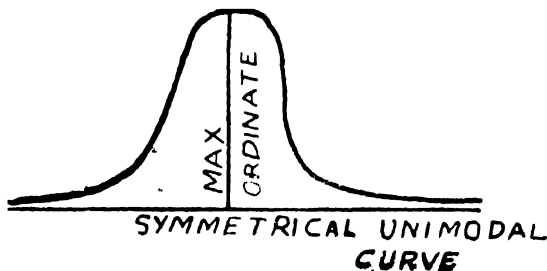


Fig. 1

go on decreasing symmetrically to zero on both the sides of the central (maximum) ordinate. A very important of this type is the normal curve.

(2) *Skewed Unimodal curves* : In this type of curves, the frequencies fall more rapidly on one side of the maximum ordinate

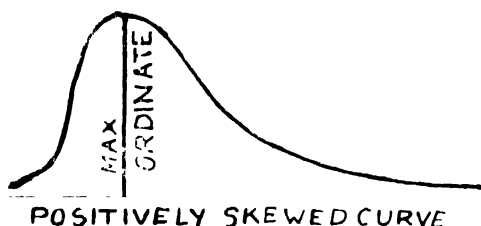


Fig. 2 (a)

than that of the other. Such curves have a tail on one side of the max. ordinate. If the tail is on the right hand side of the max. ordinate, the curve is said to be the *+vely Skewed*.

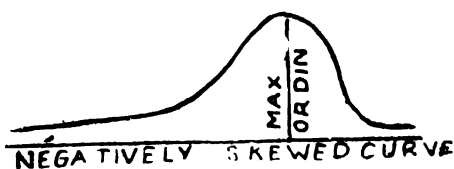


Fig. 2 (b)

When the tail is on the left hand side of the max. ordinate the curve is said *—vely Skewed curves*.

(3) *J-Shaped curves* : In this type of distributions, the maximum frequency falls on one end of the distribution; such as in

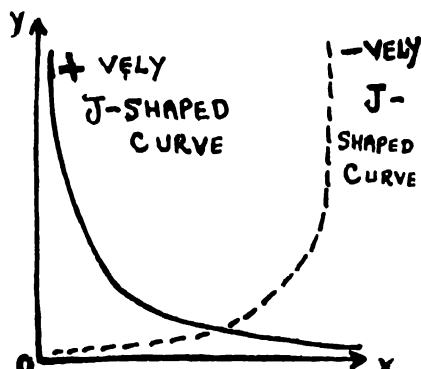


Fig. 3

the bank-balances and other income-distributions. They may be *+vely*. or *—vely J-shaped*.

(4) *U-shaped or anti-modal curves*: In this type of curves, the frequencies start from a maximum value and then fall to a minimum and again increase. These curves may be symmetrical or may not be. Such type of distributions occur in the number of unemployed persons by age-groups.

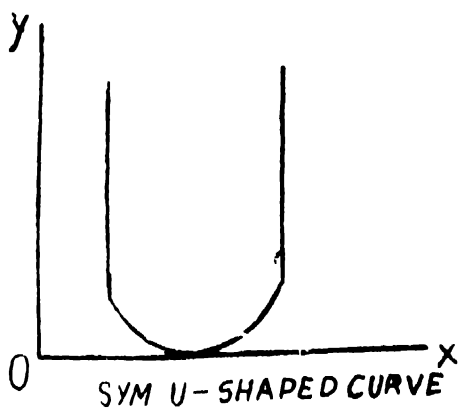


Fig. 4 (a)

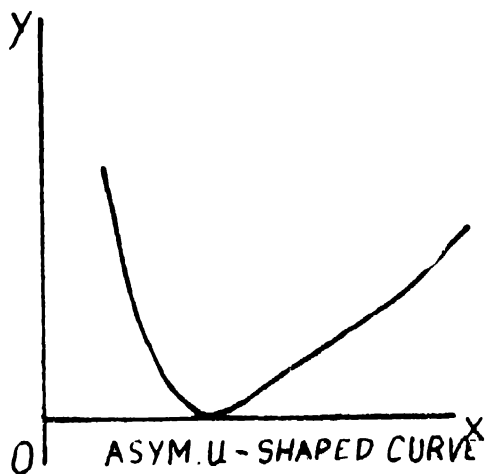


Fig. 4 (b)

EXERCISE No. (I)

(Problems on Ch. I, II, III)

Q. 1. The yield of rice in 55 equal plots of a village is given in maunds as follows

62 59 23 27 77 33 43 13 47 89 32
63 24 17 72 30 45 42 8 26 36 21
25 13 64 37 12 48 5 29 41 40 51
35 67 36 18 28 92 92 2 20 55 27
65 52 15 21 78 34 40 16 57 81 31

Prepare a frequency table taking a suitable class-interval and draw the histogram, frequency polygon and the frequency curve.

(U. P. Board, 1964)

Q. 2. (a) What is a frequency distribution? What considerations are involved in forming a frequency table from a body of observational data on a quantitative character?

(M. Sc. Ag. Agra, 1956)

(b) Make a frequency table having grades of wages with class-intervals of five annas each from the following data of daily wages received by 40 labourers in a certain factory—

5, 15, 30, 22, 30, 25, 40, 10
6 20 15 25 32 22 20 10
7 8 17 20 32 22 11 15
8 11 11 22 8 35 25 35
11 25 6 15 20 37 20 42

Q. 3. Draw the ogive of the classified data obtained from Q. No. (2) (b) and find out median & quartiles from it, graphically?

Q. 3. In the following table, the heights of 2 plants are given in centimetres at various ages—

Age (days)	Height of plant	
	A in cm.	B in cm.
40	153	156
60	167	173
80	182	180
100	187	185
120	198	195
140	201	200
160	219	219
180	220	231
200	220	231

Draw the graphs for the growth of the two plants and comment on their growth.

Q. 5. (a) Represent the following data by a suitable diagram—

Dose of nitrogen in Kgm./Hectre	Height of plants in cms. after	
	30 days	70 days
0	28	52
22	35	64
44	40	75
66	38	67
70	37	63

(b) The following table gives the total outlay on rural development proposed in the first five years plan and its break down into major items. Give a suitable diagram to represent the data—

Item	Amount (in crores of rupees)
Agr. & community development	371.43
Irrigation	178.97
Irrigation and power	276.90
Power	138.54
Transport and communication	508.10
Industry	184.04
Social Services	350.81
Rehabilitation	96.00
Miscellaneous	62.99
Total	2167.78

Q. 6. The consumer's price-index and its three most important component indexes for a country from 1935 to 1944 are given in the table below :

Year	Consumer's Price Index	Food	Clothing	Rent
1935	98.1	100.4	96.8	94.2
1936	99.1	101.3	97.6	96.4
1937	102.7	105.3	102.8	100.9
1938	100.8	97.8	102.2	104.1
1939	99.4	95.2	100.5	104.3
1940	100.2	96.6	101.7	104.6
1941	105.2	105.5	106.3	106.2
1942	116.5	123.9	124.2	108.5
1943	123.6	138.0	129.7	108.0
1944	125.5	136.1	138.8	108.2

Using graphical-method, comment particularly in respect of the periods 1937-38, 1939-40 and 1941-44.

(*M. Sc. Ag. Eco. Agra, 1965*)

Q. 7. The following data gives the marks obtained by a batch of candidates in a certain examination in Mathematics and Statistics. In which subject is the level of knowledge of the candidates higher ? Give reasons ?

Marks in Maths. : 14, 16, 16, 14, 22, 13, 15, 24, 12, 23, 14, 20, 17

Marks in Stat. : 21, 22, 18, 18, 19, 20, 17, 16, 15, 11, 12, 21, 20

Q. 8. (a) Define the median and its utility ? Calculate the median from the following data :

Class Interval	9-25	25-41	41-57	57-73	73-89	89-105	105-121	121-137	137-153	153-169
Frequency	42	48	47	40	41	21	5	2	2	2

(b) Also compute the quartiles and the quartile-deviation from the above data. *(M. Sc. Ag. Eco. Agra, 1965)*

Q. 9. The following table gives the yield of wheat from 10 equal plots :

Plot No.	1	2	3	4	5	6	7	8	9	10	Total
Yield in kgms.	60	40	50	45	60	55	65	50	65	55	45

If the area of each plot is 242 square yards, find the average yield per acre ? Also calculate the coefficient of variation (1 acre = 4840 sq. yds.). *(U. P. Board 1963)*

Q. 10. The following table gives the marks obtained by some students. Calculate the arithmetic-mean and the mode :

Marks	0-10	10-20	20-30	30-40	40-50
Frequency	3	13	18	12	5

(U. P. Board, 1963)

Q. 11. An analysis of the monthly wages paid to the workers in two firms A and B belonging to the same industry, gives the following results :

	Firm A	Firm B
No. of wage earners :	586	648
Average monthly wages :	Rs. 52.5	Rs. 47.5
Var. of the distr. of wage ;	100	121

(a) Which firm pays out the larger amount as monthly wages to the workers ?

(b) In which firm A or B, is there greater variability in individual wages ?

(c) What is the measure of average monthly wages of all the workers in the two firms A and B taken together ? *(I. A. S. 1951)*

Q. 12. Following is the distribution of marks secured by 139 candidates. Calculate the S. D. and the coefficient of variation ?

Marks (x)	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	Total
Frequency (f)	5	10	20	40	30	20	10	4	139

Q. 13. Write short notes on the following :

- (i) Classification,
- (ii) Histogram,
- (iii) Ogive,
- (iv) Frequency polygon and frequency curve,
- (v) Median and Mode.

Q. 14. The scores of two golfers for 10 rounds each are :

A : 9, 17, 14, 13, 15, 10, 11, 13, 13, 15

B : 8, 15, 11, 11, 9, 12, 11, 10, 9, 14

Which player may be regarded as more consistent ?

Q. 15. The following table gives the heights and dry weights of 17 plants. Calculate the mean and S. D. for both the characters and find which of these two varies more ?

Heights in cms. (x) :

50, 44, 54, 46, 64, 37, 17, 62, 45, 69, 44, 57, 63, 68, 45, 71, 50

Weights in gms. (y) :

44, 25, 44, 26, 41, 21, 60, 52, 15, 55, 42, 58, 37, 36, 40, 58, 36

Q. 16. In any two series, where d_1 and d_2 represent the deviations from the assumed means $A_1=50$, $A_2=154$ have $n_1=10$, $n_2=9$, $\Sigma d_1=5.8$, $\Sigma d_1^2=755.66$, $\Sigma d_2=3$, $\Sigma d_2^2=55117.0$. Calculate the coefficients of variation for the two series ?

Answers for Exercise 1.

Q. 2. (b)	<i>Class</i>	<i>Freq.</i>
	5-10	7
	10-15	6
	15-20	5
	20-25	9
	25-30	4
	30-35	4
	35-40	3
	40-45	2

The boundary line cases are kept in the classes where they are as lower limits.

Q. 7. Statistics ; (median is higher).

Q. 8. (a) Md.=53.08.

(b) $Q_1=31.916$, $Q_3=77.390$, Q. D.=22.737

Q. 9. 1090 Kgms./acre, C. V.=14.47.

Q. 10. A. M.=25.59, Mo.=24.545.

Q. 11. (a) B, (b) B, (c) 49.9 Rs.

Q. 12. S. D.=15.6, C. V.=39.2.

Q. 14. B.

Q. 15. $\bar{x}=55.65$ cms., $\bar{y}=40.59$ gms., S. D. (x)=11.71 cms.,
S. D. (y)=13.45 gms. weights vary more.

Q. 16. C. V.₁=17.1, C. V.₂=50.7.

Chapter 1V

Elementary Idea of Probability

4.1 Meaning and Definition.

The term probability is used in three different meanings—

- (1) As the subject name,
- (2) As the numerical measure of the chance that an event will happen in a single trial,
- (3) As the idea of likelihood in daily conversation in the sentences like,
 - (a) Most probably he will pass this year,
 - (b) It is very likely that his father will come back today.

The word 'Probability' used in the sense given in (2) is defined as the limit of the relative frequency as the number of trials increases indefinitely. If an event 'E' happens 'm' times in 'n' independent trials performed under the same conditions, then $\frac{m}{n}$ is known as the relative frequency. In the table given below, we are giving the results of various throws of a symmetrical coin—

No. of throws (n)	No. of times head turned up (m)	Relative frequency (R. F.)	0.5 - R.F
100	47	$\frac{47}{100} = 0.47$	0.030
200	95	$\frac{95}{200} = 0.475$	0.025
300	145	$\frac{145}{300} = 0.483$	0.017
400	195	$\frac{195}{400} = 0.487$	0.013
500	256	$\frac{256}{500} = 0.512$	0.012
600	299	$\frac{299}{600} = 0.498$	0.002

From the table we note that as the number of trials increases, the deviation of the R. F. from 0.5 decreases. Thus, it is expected that the relative frequency will be very—very near to 0.5 for some large number of throws or in other words, we can say that the

limit of the R. F. is 0.5 as the number of trials increases indefinitely. Hence 0.5 is the probability of turning up the head which means that the head will turn up 50% of the total throws of a symmetrical coin on the average in the long-run. Now, we can say that the probability is the numerical measure of the chance that in a single trial an event will happen and is defined as *the limit of the ratio of the number of happenings of an event to the total number of trials performed under the same conditions provided this limit is unique and finite*. Suppose an event 'E' happens 'm' times in 'n' independent trials then the $P(E) = \lim_{n \rightarrow \infty} \frac{m}{n}$, provided the limit is finite and unique and the trials are performed under the same set of conditions.

$P(E)$ is estimated by $\frac{m}{n}$. For example, if in a certain factory 5% of the items are found to be defective on the average during an inspection of a large number of items, then the probability that an item selected at random will be defective is $5/100 = 0.05$.

4.2 Events :

Compound Events : When an event is decomposable into a number of simple events, then it is called a *compound event*. For example, the event, the sum of the two numbers shown by the upper faces of the two dice is six in the simultaneous throw of two unbiased dice, is a compound event. Since it can be decomposed into the following simple events—

(1, 5), (2, 4), (3, 3), (4, 2), (5, 1), where the numbers in each bracket are shown by the upper faces of the 1st and 2nd die respectively.

Independent Events : The events are said to be independent when the happening of one does not affect the happenings of the others. In the reverse case, the events will be called *the dependent events*. For example, if a bag contains ten balls, and one ball is drawn from it and it is not replaced back, then a second ball is drawn from it ; the probability of the second drawing is dependent of the first and so the two draws (events) are dependent. On the other hand if the first ball is replaced, the probability of the second drawing is independent of the first and so the two drawings (events) will be independent.

Now we state without proof certain theorems which governs the probability of compound events—

Mutually Exclusive Events : The set of events is said to be

mutually exclusive when the happening of one excludes the happening of the other *i.e.* no any two events can occur simultaneously. For example the two events, the head and the tail cannot occur together in tossing a single coin (here we have ruled out the possibility of standing the coin on edge). Similarly the six possible outcomes (events) in throwing a die are *mutually exclusive*

Equally likely Events : When the probability of happening of two or more events is the same, they are called *equally likely events*. For example, the six possible outcomes of a perfectly—symmetrical die and two possible outcomes of a perfectly symmetrical coin are equally likely events. Such a perfectly symmetrical die and coin are called *unbiased*.

Exhaustive Events : The set of events is said to be exhaustive when it includes all the possible outcomes of a trial and it is certain that one of them will occur.

Compatible Events : The set of events is said to be compatible when two or more of them can happen simultaneously.

4.3 Laws of Probabilities :

(1) *Addition Theorem of probabilities :* If E_1, E_2, \dots, E_k are 'K' mutually exclusive events with their respective probabilities of happenings p_1, p_2, \dots, p_k , then the probability that any one of them will happen is $(p_1 + p_2 + \dots + p_k)$. Symbolically,

$$P(E_1 + E_2 + \dots + E_k) = \sum_{i=1}^k p_i$$

(2) *Multiplication theorem of probabilities :* If E_1, E_2, \dots, E_k are 'k' independent compatible events with p_1, p_2, \dots, p_k probabilities of their occurrences respectively, then the probability for their simultaneous occurrence is $(p_1 \cdot p_2 \cdot \dots \cdot p_k)$. Symbolically,

$$P(E_1 \cdot E_2 \cdot \dots \cdot E_k) = p_1 \cdot p_2 \cdot \dots \cdot p_k = \prod_{i=1}^k p_i$$

(3) If \bar{E} is the event opposite to E, then $P(E) + P(\bar{E}) = 1$.

Exp. (1) [a] A number is selected from each of the two sets 1, 2, 3, 4, 5, 6, and 1, 2, 3, 4, 5, 6; what is the probability that the number 1 will be selected from the first set and 3 from the second ?

[b] What is the probability that the sum of the two numbers selected one from each set is 5 ?

OR

Two dice are thrown simultaneously :—

Agricultural Statistics

[a] What is the probability that 1 will be shown by the upper face of the first die and 3 by that of second ?

[b] What is the probability that the sum of the two numbers shown by the upper faces of the two dice is 5 ?

Sol : [a] There are six numbers in this first set and in random selection for any one of them every number has an equal chance of being selected. Hence, $P(1) = \frac{1}{6}$

Similarly,

$$P(3) = \frac{1}{6}$$

The selection of (1, 3) is a compound event composed of two simple events which are independent.

Hence the probability of their simultaneous occurrence is—

$$P(1, 3) = P(1) \cdot P(3) \quad (\text{by the multiplication theorem of probabilities})$$

$$= \frac{1}{6} \times \frac{1}{6} = \frac{1}{36} \quad \text{Ans.}$$

[b] The sum of the two numbers one selected from each set would be 5 in the following 4 ways—

(1, 4), (2, 3), (3, 2), (4, 1).

Arguing in the same manner as above, we have—

$$P(1, 4) = \frac{1}{36}, P(2, 3) = \frac{1}{36}$$

$$P(3, 2) = \frac{1}{36}, \text{ and } P(4, 1) = \frac{1}{36}$$

Hence, the desired probability is—

$$P[(1, 4) + (2, 3) + (3, 2) + (4, 1)] = \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36}$$

$$= \frac{4}{36} = \frac{1}{9} \quad \text{Ans.}$$

Exp. (2) Two cards are drawn from a deck of well shuffled cards. What is the probability that both the extracted cards are aces ?

Sol : Since there are 52 cards in the deck and 4 are aces among them. In the drawing of the first card the probability that the extracted card is an ace is $\frac{4}{52}$. After the first card has been drawn, the second extracted card may be any one of the remaining 51 cards containing 3 aces. The probability of the second extracted card that it will be an ace is $\frac{3}{51}$. Hence the probability that both the extracted cards are aces, is—

$$\frac{4}{52} \times \frac{3}{51} = \frac{1}{13} \times \frac{1}{17} = \frac{1}{221} \quad \text{Ans.}$$

Exp. (3) Two cards are drawn from a full pack with replacement. What is the probability that both the extracted cards are of a specified suit?

Sol. : In the first draw, there are 52 cards in all, and 13 of them are of a specified suit. Hence the prob. that the first extracted card will belong to a specified suit is $\frac{13}{52}$. In the second draw, the total no. of cards is again 52 since the first extracted card has been replaced and also the no. of cards of a specified suit is again 13, so the prob. that the second extracted card will belong to the suit of the first extracted card is $\frac{13}{52}$. Now the probability that both the extracted cards belong to a specified suit is

$$\frac{13}{52} \times \frac{13}{52} = \frac{1}{4} \times \frac{1}{4} = \frac{1}{16}. \quad \text{Ans.}$$

Exp. (4) : An urn contains 4 white and 6 black balls.

(a) One ball is drawn at random; what is the probability that it is black?

(b) A second ball is drawn after the first (without replacement) and the colour of the first extracted ball was not noted. What is the probability that the second ball is black?

(c) What is the probability that the second extracted ball is black when the colour of the first ball was noted as black?

Sol. (a) We may imagine that the balls are numbered from 1 to 10 in such a way that the white balls bear the numbers 1, 2, 3, 4 and those of black 5, 6, 7, 8, 9 and 10. Extraction of one black ball means the selection of any one number out of 5, 6, 7, 8, 9 and 10 and it may be in six ways. Thus the prob. that a randomly selected ball is black, will be $\frac{6}{10} = 0.6$

Ans

(b) The colour of the first extracted ball may be

(i) White with prob. $\frac{4}{10}$. or (ii) Black with prob. $\frac{6}{10}$.

(i) If the colour of the first ball is white and it is not replaced back, the prob. that the second extracted ball is black, will be $\frac{6}{9}$. Hence the prob. that the first extracted ball is white and the second black when the first ball was not replaced, is—

$$\frac{4}{10} \times \frac{6}{9} = \frac{2}{5} \times \frac{2}{3} = \frac{4}{15}$$

(ii) If the colour of the first extracted ball is black and it is not replaced back, the prob. that the second extracted ball is black, will be $\frac{5}{9}$. Hence, the prob. that the first extracted ball is black and the second also black when the first ball was not replaced, is—

$$\frac{6}{10} \times \frac{5}{9} = \frac{3}{5} \times \frac{5}{9} = \frac{1}{3}.$$

The desired prob. is the sum of the probs. obtained in (i) and (ii) i.e.

$$\frac{4}{15} + \frac{1}{3} = \frac{4+5}{15} = \frac{9}{15} = 0.6. \quad \text{Ans.}$$

(c) When the first extracted ball is not replaced and is colour is black, the prob. that the second extracted ball will also be black, is— $\frac{5}{9}$. **Ans.**

Exp. (5) : A problem is given to 3 students whose chances of solving it are $\frac{1}{2}$, $\frac{1}{3}$ and $\frac{1}{4}$. What is the probability that the problem will be solved ?

Sol : If E denotes the event that the problem will be solved, then \bar{E} will be the event that it will not be solved. The problem will not be solved when each student fails to solve it.

The prob. that it will not be solved by the first student is $1 - \frac{1}{2} = \frac{1}{2}$. Similarly the prob. that it will not be solved by the second student is $1 - \frac{1}{3} = \frac{2}{3}$ and that of third is $1 - \frac{1}{4} = \frac{3}{4}$.

The prob. that none will be able to solve the problem is—

$$P(\bar{E}) = \frac{1}{2} \times \frac{2}{3} \times \frac{3}{4} = \frac{1}{4}$$

Hence, $P(E) = 1 - P(\bar{E}) = 1 - \frac{1}{4} = \frac{3}{4}$. **Ans.**

Exp. (6) : The probability that a worker selected at random from a certain factory is male is 0.6 and that of a worker is married is 0.7. Find the probability that a worker selected at random

(i) is a married male,

(ii) is a married female, and

(iii) a single female ?

Sol. If A denotes the event that a randomly selected worker is a male then \bar{A} will be the event that the worker is a female.

Similarly, if B stands for a married worker, then \overline{B} stands for a single worker. We have been given

$$P(A) = 0.6, P(B) = 0.7$$

Since we know that, $P(A) + P(\overline{A}) = 1$

or
and
or

$$P(\overline{A}) = 1 - P(A) = 0.4$$

$$P(B) + P(\overline{B}) = 1$$

$$P(\overline{B}) = 0.3$$

(i) Hence,

$$P(AB) = P(A) \cdot P(B)$$

$$= 0.6 \times 0.7 = 0.42$$

... ..

Ans.

(ii) $P(\overline{A}B) = P(\overline{A}) \cdot P(B)$

$$= 0.4 \times 0.7 = 0.28$$

... ..

Ans.

(iii) $P(\overline{A}\overline{B}) = P(\overline{A}) \cdot P(\overline{B})$

$$= 0.4 \times 0.3 = 0.12$$

... ..

Ans.

Exp. (7): In cotton F_2 segregating for leaf shape (narrow versus broad) and flower colour (yellow versus white), characters controlled by single gene pairs segregating independently and exhibiting dominance of narrow leaf and yellow flower. What is the probability of a plant selected randomly from this F_2 possessing

(i) narrow leaves,

(ii) narrow leaves and yellow flowers,

(iii) broad leaves and white flowers (*M. Sc. Ag. Agra. 1953*)

Sol. We have been given that—

(i) the plants with narrow and broad leaves are in the ratio 3 : 1,

(ii) the plants with yellow and white flower colour are in the ratio 3 : 1, and

(iii) the two characters (Shape of leaf and flower colour) are segregating independently.

The prob. that a randomly selected plant from this F_2 possessing—

(i) narrow leaves $= \frac{3}{4}$ **Ans.**

(ii) narrow leaves

and yellow flowers $= P(\text{narrow leaves}) \times P(\text{yellow flowers})$

$$= \frac{3}{4} \times \frac{3}{4} = \frac{9}{16} \text{ **Ans.**}$$

(iii) broad leaves

and white flowers $= P(\text{broad leaves}) \times P(\text{white flowers})$

$$= \frac{1}{4} \times \frac{1}{4} = \frac{1}{16} \text{ **Ans.**}$$

4.4 Binomial law of Probability:

If the probability of happening of an event in a single trial is 'p' and that of failure is 'q', then the probability that out of 'n' independent trials performed under the same conditions, the event will happen 'x' times and fail to happen (n-x) times, is given by—

$$P(x) = {}^nC_x (p^x) (q^{n-x}), \text{ where: } p+q=1 \text{ and } {}^nC_x = \frac{n!}{x! (n-x)!}.$$

The possible outcomes of the above trials are $x=0, 1, 2, \dots, n$. Thus it is a set of (n+1) events such that one of them is certain to happen, so the sum of their probabilities is unity i. e.

$$P(0) + P(1) + P(2) + \dots + P(n) = 1.$$

Exp. No. (8): An unbiased coin is tossed 3 times. What is the probability that the head will occur,

- (i) 0 time, (ii) 1 time,
(iii) 2 times and (iv) 3 times ?

Sol. Here, $n=3, p=q=\frac{1}{2}$, if $p(x)$ denotes the prob. that the head turns up 'x' times then

$$(i) \quad P(0) = {}^3C_0 \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^3 \\ = 1 \times 1 \times \frac{1}{8} = \frac{1}{8} \text{ Ans.}$$

$$(ii) \quad P(1) = {}^3C_1 \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^2 \\ = 3 \times \frac{1}{2} \times \frac{1}{4} = \frac{3}{8} \text{ Ans.}$$

$$(iii) \quad P(2) = {}^3C_2 \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^1 \\ = 3 \times \frac{1}{4} \times \frac{1}{2} = \frac{3}{8} \text{ Ans.}$$

$$\text{and (iv)} \quad P(3) = {}^3C_3 \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^0 \\ = 1 \times \frac{1}{8} \times 1 = \frac{1}{8} \text{ Ans.}$$

Exp. No. 9): Five aeroplanes fly together. What is the probability that all will reach safe to their destination? When the probability of disturbance in any one of them is 0.1?

Sol. : Here $n=5, q=0.1, p=1-q=0.9$.

If $p(x)$ denotes the prob that x planes will reach safe to their destination, then—

$$P(5) = {}^5C_5 (0.9)^5 (0.1)^{5-5} \\ = 1 \times (0.9)^5 \times 1 \\ = 0.59049 \text{ Ans.}$$

4.5 Normal law of Probability-

When the no. of trials increases indefinitely and neither 'p' nor 'q' is very small, then the limiting form of the Binomial Law is the normal law given by—

$$f(x) = \frac{1}{\sqrt{2\pi npq}} e^{-\frac{1}{2}(x-np)^2/npq}, \text{ which on putting } np=m$$

& $npq = \sigma^2$ takes the general form—

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-m)^2}{\sigma^2}}, \text{ where } m, \sigma^2 \text{ are the mean}$$

and variance of the variate x . Here x is a continuous variate which ranges from $-\infty$ to $+\infty$. The frequency curve of the normal law is of the form—

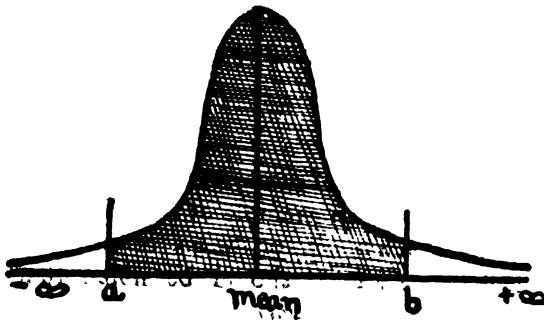


Fig. No. (1)

It is symmetrical about the mean ' m '. In fact the mean, median and the mode coincide in this case. The prob. that ' x ' will take its value between ' a ' & ' b ' is the shaded area of the above curve. For the standard normal law (with mean zero and variance unity) such probabilities are given in 'Fisher & Yates Statistical tables'. Since the normal law is found to hold to a large mass of agricultural data, hence it is used very widely in connection with the agricultural experiments especially when we deal with the large data.

Exp. No. (10) : What is probability ? For a character distributed normally with unit variance what are the probabilities of occurrence of individuals exceeding the mean value by 1 or more, 2 or more, and 3 or more units ?

(M. Sc. Ag. Agra, 1956)

Sol. : For definition see theory.

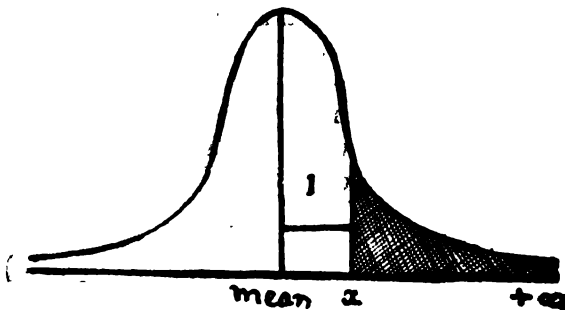


Fig No. (2)

(i) The prob. that the individual 'x' will exceed the mean m' by 1 or more is the shaded-area of the curve, which will be found out by normal tables i.e., $P\{(x-m) \geq 1\} = 0.1587$.

Similarly (ii) $P\{(x-m) \geq 2\} = 0.0228$,

and (iii) $P\{(x-m) \geq 3\} = 0.00135$.

Poisson Law of Probability :

If p or q is very small and np ($=m$ say) is a constant, then the limiting form of the *Binomial Law* is the *Poisson Law*, given by

$$P(x) = e^{-m} \frac{m^x}{x!}$$

Here x denotes the no. of successes which is a discrete variate and ranges from 0 to ∞ . An example of the same may be the no. of misprints per page in a book, or the no. of road accidents per day on a particular high way. It should be noted that the mean and the variance of this *Poisson variate* x is the same i.e.m.

Exercise No. IV

1. How will you define probability ? Give its uses.
2. What do you understand by binomial variate and normal variate ? Describe the important properties of normal probability curve and give its importance
3. Give the equation for a normal curve whose mean is zero and standard deviation is unity. What are the properties of this curve.
4. Give without proof the laws of probability.
5. What is binomial law of probability ? Using it, find the probability that all the five games will be won by a player A when the probability of missing a game is 0.1.

(Ans $p = 0.59$)

6. What is the equation for a standard normal curve. Using it, find the total number of candidates out of 1000 who obtain more than 80% marks in an examination in which the average of marks is 50 and standard deviation is 10.

(Ans 27)

7. How will you define an event. what do you mean by independent and mutually exclusive events ? Give at least two examples for each.

Chapter V

Tests of Significance

5.1 Introduction.

The aim of any statistical inquiry is to draw conclusions regarding the population characteristics (called the parameters) like mean and standard deviation, by studying the sample mean and sample s. d. etc. i.e. arriving at generalizations from a study of particular cases. In doing this, we must know the mathematical form of the population to be studied and the technique of sampling. Fortunately, the normal distribution fits a large mass of agricultural data. For the present discussion, the sample will be considered a simple random sample throughout this chapter i.e. the theory is based upon the assumption of normality of the parent population and simple random sampling. In drawing the conclusions regarding the population from the sample observations, we are faced with two types of problems. These are—

- (1) to estimate the unknown parameters of the parent population by some suitable function of the sample observations (statistics),
- (2) to compare these calculated values and to find how far the difference between the two values can be attributed to the fluctuations of simple random sampling (chance).

The former problem is called the problem of 'Estimation' while the latter as the problem of 'testing of hypothesis.' A *statistical hypothesis* is a statement which specifies the value of one or more parameters or gives the relation between the two or more parameters. Any statistical hypothesis under test is called a *null hypothesis* and it is so set up that the difference between the two values to be compared is due to chance alone. In applying a test of significance, we calculate the probability of occurrence a difference equal to or more than the observed difference between the two values to be compared under the assumption that the null hypothesis is true. In advance of the experiment, we fix up the value of such a probability which is used as a line of demarcation between the rejection and the acceptance of the null hypothesis, called the *level of significance*. These levels of significance to test the hypotheses are generally 1% and 5% in practice. If the calculated probability is less than the

level of significance, the null hypothesis will be rejected otherwise it will be accepted. If the probability of getting a difference equal to or more than the observed difference between the two values to be compared is less than .05, then the difference is said to be significant at 5 percent level of significance and when this probability is greater than or equal to .05, the difference is not significant at the 5 percent level of significance. At the 5 percent level of significance, a true hypothesis will be rejected 5 times in 100 on the average in the long run.

5.2 Sampling Distribution. If we take a number of samples from the specified population and calculate some statistic as mean or the s. d., the obtained values will be a series of different values. These values under certain consideration may be grouped in the form of a frequency distribution. If the no. of samples drawn be larger and larger, the frequency distribution tends to a continuous distribution which very often is a normal distribution. *This distribution is called the Sampling Distribution.* The test of significance is carried out by calculating the probability of observing the difference between the two values to be compared. This value can be found only when we have an idea of the sampling distribution of the statistic to be used in testing the significance. We give below the sampling distributions of a few statistics with the standard-errors of the differences between the two values to be compared.

(1) $z = \frac{(\bar{x} - \mu)}{\sigma/\sqrt{n}}$ follows a standard normal distribution,

where \bar{x} is sample mean, $\mu \rightarrow$ the population mean,

$\sigma \rightarrow$ the s.d. of the population and $n \rightarrow$ the no. of observations in the sample.

S. E. of $(\bar{x} - \mu) = \sigma/\sqrt{n}$.

(2) $t = \frac{(\bar{x} - \mu)}{s/\sqrt{n}}$ follows the 't' distribution with $(n-1)$ d. f.,

$n \leq 30$, where 's' is the s. d. of the sample given by

$s^2 = \frac{\sum (x - \bar{x})^2}{n-1}$, which is an unbiased estimate of σ .

(Due to 'Student' W. S. Gosset)

(3) $z = \frac{(\bar{x} - \mu)}{s/\sqrt{n}}$ follows a standard normal distribution when

$n > 30$.
S. E. of $(\bar{x} - \mu) = s/\sqrt{n}$.

(4) $z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})}}$ follows a standard normal

distribution, where \bar{x}_1 and \bar{x}_2 are the means of 1st and 2nd samples respectively of the sizes n_1, n_2 and σ_1^2, σ_2^2 are the variances of the two parent populations.

$$\text{S. E. of } (\bar{x}_1 - \bar{x}_2) = \sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)}.$$

$$(5) \quad t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{S \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \text{ follows a 't' distribution with } (n_1 + n_2 - 2) \text{ d.f.,}$$

where s^2 is the pooled estimate of population variance given by

$$s^2 = \frac{\Sigma(x_1 - \bar{x}_1)^2 + (x_2 - \bar{x}_2)^2}{n_1 + n_2 - 2}, \text{ where } n_1 + n_2 - 2 \leq 30.$$

$$\text{S. E. of } (\bar{x}_1 - \bar{x}_2) = S \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

(Due to Prof. R A. Fisher)

$$(6) \quad z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}} \text{ follows a standard normal}$$

distribution, where s_1^2 and s_2^2 are the sample variances given by

$$s_1^2 = \frac{\Sigma(x_1 - \bar{x}_1)^2}{n_1 - 1} \text{ and } s_2^2 = \frac{\Sigma(x_2 - \bar{x}_2)^2}{n_2 - 1}.$$

$$\text{S. E. of } (\bar{x}_1 - \bar{x}_2) = \sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}, \text{ where } n_1 \text{ and } n_2 \text{ are large.}$$

$$(7) \quad F = \frac{s_1^2}{s_2^2} \text{ (if } s_1^2 > s_2^2 \text{) follows 'F' distribution with}$$

$v_1 = (n_1 - 1)$ and $v_2 = (n_2 - 1)$ d.f. (Due to prof. G.W. Snedecor)

$$(8) \quad \chi^2 = \Sigma(O - E)^2 / E \text{ follows a } \chi^2 \text{ distribution with } v. \text{ d. f.,}$$

where v is the no. of cells whose frequencies can be determined independently Also, O and E are observed and hypothetical (expected) frequencies respectively

(Due to K. Pearson)

5.3 Degrees of freedom :

The number of independent variables used in the computation of a statistic is called its degrees of freedom. If the total number of variables is n and they are imposed under k restrictions then the degrees of freedom is $(n - k)$.

5.4 Testing the significance of difference of two means :

It consists of dividing the difference of the two means by the standard error of the difference and then finding the probability of observing this ratio. The computations of the s.e.s and the above mentioned probability depend upon the following informations :—

(i) Whether the s.d. of the parent population is known or unknown.

(ii) Whether the sample sizes are small or large.

The samples of sizes greater than 30 are considered as large samples. The theory will be discussed for the following situations;—

(i) When ' σ ' is known, whatever be the sample size (large or small).

(ii) When ' σ ' is unknown and the sample is large.

(iii) When ' σ ', is unknown and the sample is small.

5 4.1. When ' σ ' is known and sample is of any size: We shall discuss here two types of problems :—

(a) To test whether the observed sample mean (\bar{x}) is significantly different from a 'specified population (universe)-mean (μ) i.e. to test whether a given sample has been taken from a population of specified mean.

(b) To test the significance of the difference viz. ($\bar{x}_1 - \bar{x}_2$) between the two sample means \bar{x}_1 and \bar{x}_2 say.

Case (a): Let us suppose that $x_1, x_2 \dots x_n$ constitute a simple random sample of size ' n ' from a normal population with mean μ and standard deviation σ .

To test the *null hypothesis* (denoted by N. H. or H_0) 'that the sample has been taken from a specified population' (to test the significance of the difference, $\bar{x} - \mu$), we compute the statistic

$$z = \frac{|\bar{x} - \mu|}{\sigma/\sqrt{n}}.$$

The sampling distribution of this statistic ' z ' is normal with mean zero and variance unity. Hence the table of standard normal variate can be used to determine the probability of observing the value of z . If this probability is greater than 0.05 the hypothesis (H_0) will not be rejected at 5 percent level of significance. The same can be done by computing z against 1.96. If $z \geq 1.96$, the probability of observing z is less than or equal to 0.05 leading to the rejection of the null hypothesis and if $z < 1.96$, the probability of observing z is greater than 0.05 leading to the acceptance of the null hypothesis. In short, the test is carried out as follows :—

$$\text{compute } z = \frac{|\bar{x} - \mu|}{\sigma/\sqrt{n}}$$

If $z \geq 1.96$, reject the N. H.

but if $z < 1.96$, there is no evidence against the Hyp. at 5% level of significance.

The above test is based upon the following assumptions:—

- (1) The parent population is normal.
- (2) The sample is a simple random sample.
- (3) The s. d. of the population is known.

Case (b) : Let us suppose that x_1, x_2, \dots, x_{n_1} is a s. r. s. of size n_1 from a normal population with mean μ_1 & s.d. σ_1 and a second sample of size n_2 is also a s.r.s. drawn from another normal population with mean μ_2 & s.d. σ_2 . To test the hypothesis that the two samples have been taken from the two different populations with the same mean i.e. $\mu_1 = \mu_2$, we compute the statistic

$$\text{or } z = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)}}$$

If $z \geq 1.96$, reject the N. H.

but if $z < 1.96$, there is no evidence against the hyp at 5 percent level of significance.

If the test is to be carried out at 1 percent level of significance then z is compared against 2.58 instead of 1.96.

The assumptions involved in the above test are:—

- (1) The two parent populations are normal.
- (2) The samples are simple random samples.
- (3) The two samples are independent.
- (4) The s. ds. of the two populations are known.

Exp. (1) : [a] A random sample of 25 items is drawn from a normal population with mean 5.15 and variance 4.0. If the sample mean is 6.65, can the sample be regarded as drawn from the specified population?

[b] A simple random sample of 9 is drawn from a normal population with mean μ_1 (unknown) and s.d. 4.5. Another simple random sample of size 10 is drawn from a normal population with mean μ_2 (unknown) and s.d. 4.66. Test, whether the means of the two populations are equal? Given that sample means are 68.0 and 69.2 respectively.

Sol. (a) : H_0 :—The sample has been taken from the specified population i.e. $\mu = 5.15$ & $\sigma^2 = 4$.

Now we compute the statistic,

$$\begin{aligned} z &= \frac{|\bar{x} - \mu|}{\sigma/\sqrt{n}} \\ &= \frac{|6.65 - 5.15|}{2/\sqrt{25}} = \frac{1.50}{2} \times 5 = \frac{7.5}{2} \end{aligned}$$

$\therefore z = 3.75$ i.e. > 1.96 .

Thus we reject the N. H. at 5 percent level,

Conclusion : Therefore, we conclude that the sample has not been drawn from the specified population,

[b] $H_0 : -\mu_1 = \mu_2$.

Here we compute the statistic, $z = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)}}$

$$\text{or } z = \frac{|68.0 - 69.2|}{\sqrt{\left\{\frac{(4.5)^2}{9} + \frac{(4.66)^2}{10}\right\}}} = \frac{1.20}{\sqrt{(4.42)}} = \frac{1.20}{2.1}$$

$$\therefore z = 0.57 \text{ i. e. } < 1.96.$$

Thus we have no evidence against the hyp at 5 % level.

Result : Therefore, we conclude that the two samples have been drawn from the two different populations with the same mean.

Exp. (2) : The means of simple samples of 1000 and 2000 are 67.50 and 68.0 inches respectively. Can the samples be regarded as drawn from the same population of standard deviation 2.25 inches ?
(M. Sc. Agra, 1954)

Sol. H_0 : The samples have been taken from the same universe i.e. $\mu_1 = \mu_2$ and $\sigma_1 = \sigma_2 = 2.25$.

Now we compute the statistic

$$z = \frac{|\bar{x}_1 - \bar{x}_2|}{\sigma \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad (\text{Since } \sigma_1 = \sigma_2 = \sigma \text{ is known})$$

$$= \frac{|67.50 - 68.0|}{2.25 \sqrt{\left(\frac{1}{1000} + \frac{1}{2000}\right)}} = \frac{0.50}{0.97} = 5.154.$$

So $z > 1.96$ and we reject the N.H. at 5% level.

Conclusion : We therefore conclude that the samples have not been drawn from the same population.

5.4.2 When 'σ' is the unknown and sample is large :

Case (a) : If it is desired to test the hyp.—whether a given sample has been taken from a population of specified mean (μ); we compute the statistic

$$z = \frac{|\bar{x} - \mu|}{s/\sqrt{n}}, \text{ where 's' is the unbiased estimate of } \sigma \text{ given}$$

$$\text{by } s^2 = \frac{(x - \bar{x})^2}{n-1}.$$

If $z \geq 1.96$, reject the N. H.

but if $z < 1.96$, accept the N.H. at 5 percent level of significance.

The assumptions involved in the above test are—

- (1) The parent population is normal.
- (2) The sample is a s. r. s.
- (3) The s. d. of the population is unknown.
- (4) The size of the sample is large.

Case (b) : If we want to test the hyp.—that the two samples have been taken from two different populations with the same mean *i. e.* $\mu_1 = \mu_2$, we compute the statistic

$$z = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}} \text{, where } s_1 \text{ \& } s_2 \text{ are the two unbiased estimates of the population s. ds. } \sigma_1, \sigma_2 \text{ respectively.}$$

If $z \geq 1.96$, reject the N. H.

but if $z < 1.96$, accept the N.H. at 5 percent level of significance.

The assumptions of the test are—

- (1) The parent populations are normal.
- (2) The two simple random samples are independent.
- (3) The s. ds. of the populations are unknown.
- (4) The sample sizes are large.

Exp. (3) : A rope manufacturer adopts a new process if the mean breaking strength of the 400 ropes was found to be 125 lbs. with its s. e. as 10.5. The mean breaking strength of the ropes manufactured by the old process was 105 lbs. Is the new process superior to the old one ?

Sol. H_0 : The new process of manufacturing the ropes is not superior to the old one *i.e.* $\mu = 105$ Lbs.

We are given— $\bar{x} = 125$ Lbs., s.e. of $\bar{x} = s/\sqrt{n} = 10.5$ Lbs.
 $\mu = 105$ Lbs., and $n = 400$ (large).

Thus we compute the statistic,

$$z = \frac{|\bar{x} - \mu|}{s/\sqrt{n}} = \frac{|125 - 105|}{10.5} = \frac{20}{10.5} = 1.904.$$

So $z < 1.96$ and we have no evidence against the hyp. at 5% level of significance.

Conclusion : We, therefore, conclude that the new process is not superior to the old one.

Exp. (4) : A sample of heights of 6400 soldiers has a mean of 67.85 inches and a s. d. of 2.56 inches while a simple sample of heights of 1600 sailors has a mean of 68.55 inches and a standard deviation of 2.52 inches. Do the data indicate that the sailors are on the average taller than soldiers ? (B. Sc. Agra, 1955)

Sol. H_0 : The ors are not on the average taller than soldiers *i.e.* $\mu_1 = \mu_2$.

Here we compute the statistic,

$$z = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}} = \frac{|67.85 - 68.55|}{\sqrt{\left\{\frac{(2.56)^2}{6400} + \frac{(2.52)^2}{1600}\right\}}} = \frac{0.7}{0.07} = 10.$$

So $z > 1.96$ and the hyp. is rejected at 5 percent level of significance.

Result : Hence the data indicate that the sailors are on the average taller than the soldiers.

Exp. (5) : The mean staple length determined by taking 100 samples from each of two lots of cotton was as follows—

lot A : 25.4 ± 0.8 m. m.

lot B : 26.8 ± 0.7 m. m.

Do the lots differ significantly in their mean staple lengths ?

(M. Sc. Ag. Agra, 1958)

Sol. Ho : The mean staple lengths do not differ significantly i. e. $\mu_1 = \mu_2$.

Given that $\bar{x}_1 = 25.4$ m. m., s. e. of $(\bar{x}_1) = 0.8$ m. m.

$\bar{x}_2 = 26.8$ m. m., s. e. of $(\bar{x}_2) = 0.7$ m. m.

Since the samples are large, so we compute the statistic

$$z = \frac{|\bar{x}_1 - \bar{x}_2|}{\text{s.e. of } (\bar{x}_1 - \bar{x}_2)} = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{[(\text{s.e. of } \bar{x}_1)^2 + (\text{s.e. of } \bar{x}_2)^2]}}$$

because $\text{s.e. of } (\bar{x}_1 - \bar{x}_2) = \sqrt{[\text{s.e. of } \bar{x}_1]^2 + [\text{s.e. of } \bar{x}_2]^2}$.

$$\text{or } z = \frac{|25.4 - 26.8|}{\sqrt{[0.8]^2 + [0.7]^2}} = \frac{1.4}{1.06} = 1.32.$$

So $z < 1.96$ and the hyp. is accepted at 5 percent level of significance.

Result : Therefore, we arrive at the conclusion that the mean staple-lengths do not differ significantly.

5.4.3 When 'σ' is unknown and sample is small :

Case (a) : To test the hyp.—that the sample has been taken from a normal population with specified mean (μ) and unknown s. d.; we compute the statistic 'Student-t' as

$$t = \frac{|\bar{x} - \mu|}{s/\sqrt{n}}, \text{ where } s^2 = \frac{\sum(x - \bar{x})^2}{n-1}.$$

If $t \geq t_{0.05}(n-1)$, reject the hypothesis,

but if $t < t_{0.05}(n-1)$, there is no evidence against the hyp. at 5% level. Here $t_{0.05}(n-1)$ stands for the tabulated value of 't' at 5 percent level of significance for $(n-1)$ d. f.

assumptions :

- (1) The parent universe is normal.
- (2) The sample is a simple random.
- (3) The s. d. of the population is unknown.
- (4) The size of the sample is small.

Case (b) : If we want to test the hyp.—that the two samples have been drawn from the same population, we compute the

statistic 'Fisher t' as $t = \frac{|\bar{x}_1 - \bar{x}_2|}{s \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$, where s is the pooled

estimate of the population s. d. given by $s^2 = \frac{\Sigma(x_1 - \bar{x}_1)^2 + \Sigma(x_2 - \bar{x}_2)^2}{n_1 + n_2 - 2}$.

If $t \geq t_{0.05}(n_1 + n_2 - 2)$, reject the N. H.

but if $t < t_{0.05}(n_1 + n_2 - 2)$, there is no evidence against the hyp. at 5% level of significance.

The test has the following assumptions—

- (1) The populations are normal.
- (2) The samples are simple random and independent.
- (3) The s. d. of the two populations are the same but unknown.
- (4) The sizes of the samples are small.

Case (c) : We have discussed above in case (b) that "Fisher-t test" is used only when the two samples are independent. But if the two samples give an indication of (+)ve correlation i.e. the observations are paired, then the test is carried out by computing the statistic "Paired t" as

$t = \frac{|\bar{d}|}{s/\sqrt{n}}$, where 'd' stands for the difference between the paired

values (x y) i.e. $d = (x - y)$ say, $\bar{d} = \Sigma d/n$ and $s^2 = \Sigma(d - \bar{d})^2/(n - 1)$.

If $t \geq t_{0.05}(n - 1)$, reject the N. H.

but if $t < t_{0.05}(n - 1)$, there is no evidence against the hyp. at 5% level of significance.

Assumptions :

- (1) The parent population is a bivariate normal.
- (2) The two samples show (+)ve correlation.
- (3) The s. ds of x and y are unknown.
- (4) The samples are small.

Case (d) : In the cases (b) & (c) discussed above for two sample problems, we have noted that the population-variances were unknown but they were supposed to be the same. Thus for small samples there may arise one more situation where the population-variance being unknown may not be the same. In such a situation for two small and independent samples 'Fisher's-t-test' cannot be applied. The exact test for this purpose was first developed by *Behrens* and later on by *Fisher*, and so it is called 'Fisher-Behrens-t-test'. The tables for the applications of this test

are prepared by Dr. P. V. Sukhatme. But in using these tables, the interpolation is used a no. of times and so this test is a complicated one ?

In practice, we use an approximate test namely '*Cochran and Cox-t-test*' which is usually carried out after justifying the fact that $\sigma_1^2 \neq \sigma_2^2$. Thus in testing the hyp. $\mu_1 = \mu_2$, first we should test the hyp. $\sigma_1^2 = \sigma_2^2$ through the F-test. If F-test gives no evidence against the hyp. (say at 5%), then we should apply *Fisher's-t-test* otherwise *Cochran and Cox-t-test*. For Cochran and Cox-t-test,

we compute the statistic $t = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}}$ and compare it against

$$t' = \frac{\frac{t_1 s_1^2}{n_1} + \frac{t_2 s_2^2}{n_2}}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}. \text{ Here } t' \text{ is a weighted mean of the tabulated}$$

values $t_1 = t_{0.05}(n_1-1)$ and $t_2 = t_{0.05}(n_2-1)$ with weights as s_1^2/n_1 and s_2^2/n_2 respectively, where s_1^2, s_2^2 have their usual meanings.

If $t \geq t'$, we reject the N.H. at 5% level ;

but if $t < t'$, there is no evidence against the hyp. at 5% level.

The assumptions involved in the above test are—

- (1) The samples are small.
- (2) The two samples are simple random and independent.
- (3) The parent populations are normal.
- (4) The s.d.s. of the populations are unknown and different.

Exp. 6. (a) A certain drug was administered to each of the 13 patients and it resulted in the gain of sleeping hours as follows :

—4, 5, 2, 8, —1, 3, 0, 6 —3, 1, 5, 0, 4.

Can it be concluded that the drug will in general be accompanied by an increase in the sleeping hours ?

(b) Ten individuals are chosen at random from a population and their heights are found as follows :

63, 63, 66, 67, 68, 69, 70, 70, 71, 71 inches respectively.

Test whether the mean height is 69.6" in the population.

Given that for 9 d. f. $P\{|t| \geq 2.262\} = 0.05$,

(M. Sc. Agra, 1962)

Sol. (a) H_0 : The drug does not have any increase in the sleeping hours of the patients i. e. $\mu = 0$.

Since the sample is small and the s.d. of the population is

unknown, so we compute the statistic $t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$. The computations for \bar{x} and s are shown in the following table.

	Gain x hrs.	x^2	Computations
1	-4	16	$\bar{x} = \frac{\Sigma x}{n} = \frac{26}{13} = 2 \text{ hrs.}$ and $s^2 = \frac{\Sigma(x - \bar{x})^2}{n-1}$ $= \frac{\Sigma x^2 - (\Sigma x)^2/n}{n-1}$ $= \frac{206 - (26)^2/13}{12}$ $= \frac{145}{12}$ $= 12.8333$ $\therefore s = \sqrt{12.8333}$ $= 3.58 \text{ hrs.}$
2	5	25	
3	2	4	
4	8	64	
5	-1	1	
6	3	9	
7	0	0	
8	6	36	
9	-3	9	
10	1	1	
11	5	25	
12	0	0	
13	4	16	
$n=13$	$\Sigma x=26$	$\Sigma x^2=206$	

Thus we have

$$t = \frac{|2-0|}{3.58/\sqrt{13}} = \frac{2}{3.58/3.6} = 2.01.$$

or $t = 2.01$ and $t_{.05}(12) = 2.179$.

So $t < t_{.05}(12)$ leading to the acceptance of N. H. at 5 percent.

Conclusion : Since the observed value of t is less than the tabulated value of t at 5 percent for 12 d. f, we conclude that the drug does not have any increase in the sleeping hours of the patients.

(b). **H₀ :** The sample has been taken from the population of specified mean : $\mu = 69.8$.

Here the sample is small and the s. d. of the population is unknown, so we shall compute the statistic $t = \frac{|\bar{x} - \mu|}{s/\sqrt{n}}$.

For the computations of \bar{x} and s , we prepare the following table—

Obs. No.	Height x''	Dev. $\xi = (x - A)$ $A = 67$	$\xi^2 = (x - A)^2$	Computations
1	63	-4	16	<p>Assumed mean $A = 67''$</p> <p>$\bar{x} = A + \frac{\sum \xi}{n} = 67 + \frac{8}{10} = 67.8''$</p> <p>and</p> <p>$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1} = \frac{\sum \xi^2 - (\sum \xi)^2/n}{n - 1}$</p> <p>$= \frac{88 - (8)^2/10}{9} = \frac{88 - 6.4}{9} = \frac{81.6}{9}$</p> <p>$= 9.0667.$</p> <p>$\therefore s = \sqrt{9.0667}$</p> <p>$= 3.01''.$</p>
2	63	-4	16	
3	66	-1	1	
4	67	0	0	
5	68	+1	1	
6	69	+2	4	
7	70	+3	9	
8	70	+3	9	
9	71	+4	16	
10	71	+4	16	
Totals $n = 10$	—	$\sum \xi = 8$	$\sum \xi^2 = 88$	

$$\text{Thus } t = \frac{67.8 - 69.61}{3.01/\sqrt{10}} = \frac{1.8}{3.01/3.06} = 1.89.$$

or $t = 1.89$ and $t_{0.05}(9) = 2.262$.

So $t < t_{0.05}(9)$ leading to the acceptance of hyp. at 5% level.

Conclusion : As the observed value of t is less than the value of t from the table at 5 percent level for 9 d.f.; we conclude that the sample is taken from the population of specified mean.

Exp. 7. (a) Explain the utility of student's small sample theory in biological research.

(b) In a field-experiment with two varieties of wheat grown in seven pairs of plots, the following yields were obtained—

Variety	Yield in mds./acre						
A	15	14	12	15	16	11	13
B	11	11	13	12	13	10	12

From the above data, compute the mean difference, standard deviation of the difference and ' t ' ? What is a reasonable inference about the population mean difference ? (*M. Sc. Ag; Agra 1964*)

Sol. (a) In biological research, the research worker is very often faced with the problem of testing :

- (i) Whether the observed sample mean is significantly different from some specified value ; e. g. if we are given the additional hours of sleep gained by 10 patients who were administered a certain drug and we want to know whether the drug is helpful in producing the additional sleep. This problem is simply to test whether the mean of additional hours of sleep is significantly different from zero. If this difference is significant then we say that the drug is helpful in producing the additional sleep.
- (ii) Whether the difference between the two sample means is significant; e.g. if we are given the mean no. of bacteria in colonies/plate obtainable by four slightly different methods from soil samples taken at 4 P.M. and 8 P.M. respectively and want to know whether the no. of bacteria at 8 P.M. is more than at 4 P.M., then we shall have to test the significance of the difference between the mean no. of bacteria at 4 P.M. and 8 P.M.
- (iii) Whether the correlation and regression coefficients are significantly different from some specified values.

The limitations of the biological data are as follows—

- (1) Their sizes are small.
- (2) The population s. d. (σ) is unknown.

The discovery of ' t ' distribution gave the exact tests for the problems of type (i), (ii) and (iii) mentioned above in the case of small samples. Before its discovery, the test applied to the above problems in the case of small data were the same as those for the large data which were not statistically valid. Thus the discovery of ' t ' distribution supplied an exact test applicable to small as well as to large samples and so introduced the modern era of statistics.

(6) Here we apply the *Paired-t-test* and the computations for it are made from the following table :—

Pair No.	Yield for A	Yield for B	Difference $d = (A - B)$	$d^2 = (A - B)^2$	Computations
1	15	11	+4	16	$\bar{d} = \frac{\sum d}{n} = \frac{14}{7} = 2$ and $s^2 = \frac{\sum (d - \bar{d})^2}{n - 1}$ $= \frac{\sum d^2 - (\sum d)^2/n}{n - 1}$ $= \frac{46 - (14)^2/7}{6}$ $= \frac{18}{6}$ $= 3$ $\therefore s = \sqrt{3}$ $= 1.732$
2	14	11	+3	9	
3	12	13	-1	1	
4	15	12	+3	9	
5	16	13	+3	9	
6	11	10	+1	1	
7	13	12	+1	1	
$n = 7$	—	—	$\sum d = 14$	$\sum d^2 = 46$	

H_0 : The two varieties of wheat are not significantly different from each other as regards their average yields/plot i.e. $\mu_1 = \mu_2$.

Since the two series are correlated series and the observations are paired, so we compute the paired-*t*-statistic as

$$t = \frac{|\text{mean difference}|}{\text{s.e. of difference}}$$

$$= \frac{|\bar{d}|}{s/\sqrt{n}} = \frac{|2|}{1.732/\sqrt{7}} = \frac{2}{1.732/2.646} = \frac{2}{.6545} = 3.06.$$

or $t = 3.06$ and $t_{0.05}(6) = 2.447$.

So $t > t_{0.05}(6)$ leading to the rejection of N. H. at 5 percent.

Result : Mean difference = 2.0 md./acre,

s.e. of difference = 0.6545 md./acre, and $t = 3.06$.

The reasonable inference about the population means is that they are not equal. Since the calculated value of '*t*' exceeds the tabulated value of '*t*' at 5 percent level for 6 d.f., so the two varieties of wheat are significantly different as regards their average yields/plot.

Exp. (8) For a random sample of 10 pigs fed on a diet A, the increases in weights in a certain period were :

10, 6, 16, 17, 13, 12, 8, 14, 15, 9 lbs.

For another random sample of 12 pigs fed on a diet B, the increases in weights in the same period were :

7, 13, 22, 15, 12, 14, 18, 8, 21, 23, 10, 17 lbs.

Find, if the two samples are significantly different regarding the effects of diets. Given that for d.f. (v)=20, 22, 24 the 5 percent values of t are respectively 2.09, 2.07, 2.06.

(M. Sc. Agra, 1963)

Sol. H_0 : The effects of the two diets on the average do not differ significantly i.e. $\mu_1 = \mu_2$.

Since the samples are independent, small and the s.d.s. of the populations are unknown, so we compute the statistic 'Fisher's-t' as

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{s\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \text{ where } s^2 = \frac{\Sigma(x_1 - \bar{x}_1)^2 + \Sigma(x_2 - \bar{x}_2)^2}{n_1 + n_2 - 2}.$$

The sample-means ' \bar{x}_1, \bar{x}_2 ' and the pooled estimate of population-variance ' s^2 ' may be computed in the following tabular form.

I-Sample				II-Sample				Computations
Obs. No.	Obs. x_1	$x_1 - \bar{x}_1$ $\bar{x}_1 = 12$	$(x_1 - \bar{x}_1)^2$			$x_2 - \bar{x}_2$ $\bar{x}_2 = 15$	$(x_2 - \bar{x}_2)^2$	
1	10	-2	4	1	7	-8	64	$\bar{x}_1 = \frac{\Sigma x_1}{n_1} = \frac{120}{10} = 12 \text{ Lbs.}$ $\bar{x}_2 = \frac{\Sigma x_2}{n_2} = \frac{180}{12} = 15 \text{ Lbs.}$ and $s^2 = \frac{\Sigma(x_1 - \bar{x}_1)^2 + \Sigma(x_2 - \bar{x}_2)^2}{n_1 + n_2 - 2}$ $= \frac{120 + 314}{10 + 12 - 2}$ $= \frac{434}{20}$ $= 21.7 \text{ Lbs.}$ $\therefore s = \sqrt{(21.7)}$ $= 4.65 \text{ Lbs.}$
2	6	-6	36	2	13	-2	4	
3	16	+4	16	3	22	+7	49	
4	17	+5	25	4	15	0	0	
5	13	+1	1	5	12	-3	9	
6	12	0	0	6	14	-1	1	
7	8	-4	16	7	18	+3	9	
8	14	+2	4	8	8	-7	49	
9	15	+3	9	9	21	+6	36	
10	9	-3	9	10	23	+8	64	
				11	10	-5	25	
				12	17	+2	4	
$n_1 = 10$				$n_2 = 12$				
$\Sigma x_1 = 120$				$\Sigma x_2 = 180$				
---				---				
$\Sigma(x_1 - \bar{x}_1)^2 = 120$				$\Sigma(x_2 - \bar{x}_2)^2 = 314$				

Thus we have

$$t = \frac{|12-15|}{4.65 \sqrt{(\frac{1}{10} + \frac{1}{10})}} = \frac{3}{4.65 \times .43} = \frac{3}{1.9995},$$

or $t = 1.504$ and $t_{.05}(20) = 2.09$.

So $t < t_{.05}(20)$, leading to the acceptance of H_0 at 5% level.

Conclusion : Since the observed value of t is less than the tabulated value of ' t ' at 5% level for 20 d.f., so there is no evidence against the hyp. at 5%. It clearly means that the two diets do not differ significantly as regards the average increase in the weights of the pigs.

Exp. (9) : Two operators 'A & B' carried out simultaneous measurements of the percentage of ammonia in a plant-gas on nine successive days. It is required to know whether the two differed in their average results under the following situations—

(a) Samples were taken and divided equally between the two operators for test.

(b) The two operators worked on independent samples.

A: 4, 37, 35, 43, 34, 36, 48, 33, 33

B: 13, 37, 38, 36, 47, 48, 57, 28, 42 (M A. Patna 1955)

Sol. (a). H_0 : The two operators do not differ significantly in their average results i.e. $\mu_1 = \mu_2$.

Here the observations are paired and we suspect a (+)ve correlation between the observations of the two operators. Further the samples are small and the s.d.s. of the populations are unknown. Thus we compute here the statistic 'Paired- t ' as

$$t = \frac{|\bar{d}|}{s/\sqrt{n}}, \text{ where } d = \text{difference of paired values, } \bar{d} = \frac{\sum d}{n},$$

$$\text{and } s^2 = \frac{\sum (d - \bar{d})^2}{n-1} = \frac{\sum d^2 - (\sum d)^2/n}{n-1}.$$

The mean ' \bar{d} ' and the variance ' s^2 ' of the differences of the paired values may be computed in the following tabular form.

Pair No.	I-Sample Obs. x_1	II-Sample Obs. x_2	Difference $d = x_1 - x_2$	d^2	Computations
1	4	18	-14	196	$\bar{d} = \frac{\Sigma d}{n} = \frac{-48}{9} = -5.33$ and $s^2 = \frac{\Sigma(d - \bar{d})^2}{n - 1}$ $= \frac{\Sigma d^2 - (\Sigma d)^2/n}{n - 1}$ $= \frac{754 - (-48)^2/9}{8}$ $= \frac{498}{8}$ $= 62.27$ $\therefore s = \sqrt{62.27}$ $= 7.8$
2	37	37	0	0	
3	35	38	-3	9	
4	43	36	7	49	
5	34	47	-13	169	
6	36	48	-12	144	
7	48	57	-9	81	
8	33	28	5	25	
9	33	42	-9	81	
$n=9$	—	—	$\Sigma d = -48$	$\Sigma d^2 = 754$	

Thus we have

$$t = \frac{|-5.33|}{7.8/\sqrt{9}} = \frac{5.33}{7.8/3} = \frac{5.33}{2.6}$$

or $t = 2.05$ and $t_{.05}(8) = 2.306$.

So $t < t_{.05}(8)$, leading to the acceptance of H_0 at 5% level.

Conclusion : The two operators do not differ significantly in their average results.

(b) H_0 : The two operators do not differ significantly in their average results i.e. $\mu_1 = \mu_2$.

Here the samples dealt by the two operators are independent, small and the s.d.s. of the populations are unknown. Thus we compute the statistic 'Fisher's-t' as

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{s\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \text{ where } s^2 = \frac{\Sigma(x_1 - \bar{x}_1)^2 + \Sigma(x_2 - \bar{x}_2)^2}{n_1 + n_2 - 2}$$

The sample-means ' \bar{x}_1, \bar{x}_2 ' and the pooled estimate of population variance ' s^2 ' may be computed in the following tabular-form.

I-Sample				II-Sample				Computations
Obs. No.	Obs. x_1	Dev. $\xi = (x_1 - A_1)$ $A_1 = 34$	ξ^2	Obs. No.	Obs. x_2	Dev. $\eta = (x_2 - A_2)$ $A_2 = 39$	η^2	
1	4	-30	900	1	18	-21	441	Let $A_1 = 34, A_2 = 39$. $\bar{x}_1 = A_1 + \frac{\sum \xi}{n_1}$ $= 34 + \frac{(-3)}{9} = 33.67,$ $\bar{x}_2 = A_2 + \frac{\sum \eta}{n_2}$ $= 39 + 0 = 39,$ and $s^2 = \frac{\sum \xi^2 - (\sum \xi)^2/n_1 + \sum \eta^2 - (\sum \eta)^2/n_2}{n_1 + n_2 - 2}$ $= \frac{1193 - (-3)^2/9 + 1054 - 0}{9 + 9 - 2}$ $= \frac{1193 - 1 + 1054}{16}$ $= \frac{2246}{16}$ $= 140.375$ $\therefore s = \sqrt{(140.375)}$ $= 11.85$
2	37	3	9	2	37	-2	4	
3	35	1	1	3	38	-1	1	
4	43	9	81	4	36	-3	9	
5	34	0	0	5	47	8	64	
6	36	2	4	6	48	9	81	
7	48	14	196	7	57	18	324	
8	33	-1	1	8	28	-11	121	
9	33	-1	1	9	42	3	9	
$n_1 = 9$	—	$\sum \xi = -3$	$\sum \xi^2 = 1193$	$n_2 = 9$	—	$\sum \eta = 0$	$\sum \eta^2 = 1054$	

Thus we have

$$t = \frac{|33.67 - 39|}{11.85 \sqrt{(\frac{1}{9} + \frac{1}{9})}} = \frac{5.33}{11.85 \times .47} = \frac{5.33}{5.57}$$

or $t = 0.95$ and $t_{.05} (16) = 2.12$.

So, $t < t_{.05} (16)$, leading to the acceptance of H_0 at 5% level.

Conclusion : The two operators do not differ significantly as regards their average measurements.

Exp. (10) : The mean of 12 observations is 5.885 with a s.e. of 0.0094. The mean of 20 observations by a different method is 5.855 with a s.e. of 0.0038. Are these means significantly different ?

Sol. H_0 : *The two means do not differ significantly i.e. $\mu_1 = \mu_2$.*
Given that :

$$n_1=12, \quad n_2=20, \quad \bar{x}_1=5.885, \quad \bar{x}_2=5.855,$$

$$\text{s. e. of } (\bar{x}_1) = \frac{s_1}{\sqrt{(n_1)}} = 0.0094,$$

$$\text{and s. e. of } (\bar{x}_2) = \frac{s_2}{\sqrt{(n_2)}} = 0.0038.$$

Here the samples are small, independent and the s.d.s. of the populations are unknown. Thus we shall use either Fisher's-t or Cochran and Cox-t-test for the purpose.

In order to decide which of these two should be applied here, we shall first test the equality of variances ($\sigma_1^2 = \sigma_2^2$) by F-test. Thus we compute,

$$F = \frac{s_1^2}{s_2^2} = \frac{12 \times (0.0094)^2}{20 \times (0.0038)^2} = \frac{26508}{7220} = 3.67$$

$$\text{or } F = 3.67 \text{ and } F_{.05}(11, 19) = 2.67.$$

$\therefore F > F_{.05}(11, 19)$, leading to $\sigma_1^2 \neq \sigma_2^2$. Hence we should compute here the Cochran and Cox-t statistic, given by

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}} \\ = \frac{|5.885 - 5.855|}{\sqrt{(0.00008836 + 0.0001444)}} = \frac{.03}{\sqrt{(0.001028)}} = \frac{.03}{.01} = 3.$$

Now we shall calculate t' as

$$t' = \frac{t_1 \frac{s_1^2}{n_1} + t_2 \frac{s_2^2}{n_2}}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \quad \text{where } t_1 = t_{.05}(11) = 2.201 \\ \text{and } t_2 = t_{.05}(19) = 2.093.$$

$$= \frac{2.201 \times 0.00008836 + 2.093 \times 0.0001444}{0.00008836 + 0.0001444} \\ = \frac{.0002247}{.0001028} = \frac{2247}{1028}$$

or $t' = 2.18$.

Now we see that $t > t'$, hence it leads to the rejection of N. H. at 5 percent.

Conclusion : The two sample-means differ significantly at 5 percent level of significance.

5.5 Testing the equality of two variances :

The statistic *Snedecor's F* is used to test the equality of two variances i.e. $\sigma_1^2 = \sigma_2^2$. The test consists of computing the ratio

$$F = \frac{s_1^2}{s_2^2} \text{ (provided } s_1^2 > s_2^2 \text{)}.$$

If $F \geq F_{.05} (n_1 - 1, n_2 - 1)$, reject the N. H. at 5%,
but if $F < F_{.05} (n_1 - 1, n_2 - 1)$, accept the N. H. at 5% level.

Here $F_{.05} (n_1 - 1, n_2 - 1)$ is the tabulated value of 'F' at 5 percent level for $(n_1 - 1)$ and $(n_2 - 1)$ d.f.

This test is based on the following assumptions :—

(1) The parent populations are normal.

(2) The samples are simple random and independent.

Exp. (11). Two random samples drawn from two normal populations are—

I : 20 16 26 27 23 22 18 24 25 19

II : 27 33 42 35 32 34 28 41 43 30 37.

Obtain the estimates of the variances of the populations and test whether the two populations have the same variances ?

(I. A. S. 1955)

Sol. (Ho) : The two samples have been taken from the populations of equal variances i.e. $\sigma_1^2 = \sigma_2^2$.

Here we compute the statistic $F = s_1^2/s_2^2$ (provided $s_1^2 > s_2^2$), where s_1^2, s_2^2 are the *estimates* of the variances of I, II populations respectively and they are computed from the following table.

I-Sample			II Sample			Computations
Obs. No.	Obs. x_1	Dev. $(x_1 - \bar{x}_1)$ $\bar{x}_1 = 22$	Obs. No.	Obs. x_2	Dev. $(x_2 - \bar{x}_2)$ $\bar{x}_2 = 35$	
1	20	2	4	127	-8	64
2	16	6	36	23	-2	4
3	26	4	16	342	7	49
4	27	5	25	435	0	0
5	23	1	1	532	-3	9
6	22	0	0	634	-1	1
7	18	-4	16	738	3	9
8	24	2	4	82	-7	49
9	25	3	9	941	6	36
10	19	-3	9	1043	8	4
				1130	-5	25
				1237	2	4
$n_1 = 10$ $\Sigma x_1 = 220$			$n_2 = 12$ $\Sigma x_2 = 420$			$\Sigma(x_2 - \bar{x}_2)^2 = 314$

Now we get

$$\bar{x}_1 = \frac{\Sigma x_1}{n_1} = \frac{220}{10} = 22,$$

$$\bar{x}_2 = \frac{\Sigma x_2}{n_2} = \frac{420}{12} = 35.$$

Also,

$$s_1^2 = \frac{\Sigma(x_1 - \bar{x}_1)^2}{n_1 - 1} = \frac{120}{9} = 13.33$$

and

$$s_2^2 = \frac{\Sigma(x_2 - \bar{x}_2)^2}{n_2 - 1} = \frac{314}{11} = 28.55.$$

Thus we compute the statistic

$$F = \frac{s_2^2}{s_1^2} \text{ (since } s_2^2 > s_1^2 \text{)}$$

$$= \frac{28.55}{13.33} = 2.14.$$

or $F = 2.14$ and $F_{.05}(11, 9) = 3.1$.

So $F < F_{.05}(11, 9)$ leading to the acceptance of H_0 at 5% level.

Conclusion : Since the calculated value of F is less than the tabulated value of F at 5% level for 11 and 9 d. f., so there is no evidence against the hyp. at 5%. Therefore, we conclude that the samples have been taken from the populations of equal variances.

Exp. (12). Show, how you would use the student's t-test and fisher's z-test to decide whether the two sets of observations—

17 27 18 25 27 29 27 23 17
and 16 16 20 16 20 17 15 21

indicate samples
drawn from the same universe ?

(M. Sc. Agra, 1949)

Sol. (H_0) : The two samples have been taken from the same universe i. e. $\sigma_1^2 = \sigma_2^2$ and $\mu_1 = \mu_2$.

Here we first compute the statistic Fisher's $z = \frac{1}{2} \log_e F$ for testing the hyp. $H_{01} : \sigma_1^2 = \sigma_2^2$. If it leads to the acceptance of H_{01} at 5% say, then only we need to proceed to student's t for testing the hyp. $H_{02} : \mu_1 = \mu_2$. The computations for Fisher's z and Student's t -tests are made from the following table.

I-Sample				II-Sample				Computations
Obs. No.	Obs. x_1	Dev.	$\xi = (x_1 - A_1)$ $A_1 = 17$	Obs. No.	Obs. x_2	Dev.	$\eta = (x_2 - A_2)$ $A_2 = 20$	
				$\xi^2 = (x_1 - A_1)^2$				
1	17	0	0	1	16	-4	16	<p>Now we get</p> $\bar{x}_1 = A + \frac{\sum \xi}{n_1} = 17 + \frac{57}{9} = 23.33,$ $\bar{x}_2 = A_2 + \frac{\sum \eta}{n_2} = 20 + \frac{-19}{8} = 17.63$ <p>Also,</p> $s_1^2 = \frac{\sum (x_1 - \bar{x}_1)^2}{n_1 - 1} = \frac{\sum \xi^2 - (\sum \xi)^2/n_1}{n_1 - 1}$ $= \frac{545 - (57)^2/9}{8} = \frac{184}{8} = 23,$ <p>and,</p> $s_2^2 = \frac{\sum (x_2 - \bar{x}_2)^2}{n_2 - 1} = \frac{\sum \eta^2 - (\sum \eta)^2/n_2}{n_2 - 1}$ $= \frac{83 - (-19)^2/8}{7} = \frac{37.875}{7} = 5.42$
2	27	10	100	2	16	-4	16	
3	18	1	1	3	20	0	0	
4	25	8	64	4	16	-4	16	
5	27	10	100	5	20	0	0	
6	29	12	144	6	17	-3	9	
7	27	10	100	7	15	-5	25	
8	23	6	36	8	21	1	1	
9	17	0	0					
$n_1 = 9$				$n_2 = 8$				
		$\sum \xi = 57$				$\sum \eta = -19$		
		$\sum \xi^2 = 545$				$\sum \eta^2 = 83$		

Thus to test the hyp. $H_{01} : \sigma_1^2 = \sigma_2^2$, we compute Fisher's z as

$$z = \frac{1}{2} \log_e F = 2.3026 \times \frac{1}{2} \log_{10} F = 1.1513 \log_{10} \frac{s_1^2}{s_2^2} \text{ (if } s_1^2 > s_2^2 \text{)},$$

$$= 1.1513 \log_{10} \frac{23}{5.42} = 1.1513 \cdot \log_{10} 4.25 = 1.1513 \times 0.6284 = .72.$$

or $z = 0.72$ and $z_{0.05}(8,7) = 0.96$.

So $z < z_{0.05}(8,7)$ leading to the acceptance of H_{01} at 5% level.

We, therefore, conclude that the samples have been taken from the populations of equal variances.

Now we need to test the hyp. $H_{02} : \mu_1 = \mu_2$ and find ourselves in the position to proceed for the computation of Student's-t as follows.

$$\begin{aligned}
 t &= \frac{|\bar{x}_1 - \bar{x}_2|}{s \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \quad \text{where } s^2 = \frac{\Sigma(x_1 - \bar{x}_1)^2 + \Sigma(x_2 - \bar{x}_2)^2}{n_1 + n_2 - 2} \\
 &= \frac{184 + 37.875}{9 + 8 - 2} = \frac{221.875}{15} \\
 &= 14.7917, \therefore s = 3.84 \\
 &= \frac{|23.33 - 17.63|}{3.84 \sqrt{\left(\frac{1}{9} + \frac{1}{8}\right)}} \\
 &= \frac{5.7}{3.84 \times .48} = \frac{5.7}{1.84} = 3.16.
 \end{aligned}$$

or $t = 3.16$ and $t_{0.05}(15) = 2.13$.

So $t > t_{0.05}(15)$ leading to the rejection of H_{02} at 5% level.

Here we conclude that the samples have not been taken from the populations of equal means.

Conclusion : Therefore, we arrive finally at the result that the samples have not been taken from the same universe.

5.6 χ^2 -Test :

If the observed data is classified in the form of a frequency distribution and we want to test whether the observed data is in agreement with certain theory or hypothesis, we compute the χ^2 -test as : $\chi^2 = \Sigma(O - E)^2 / E$, where O, E stand for observed and expected or theoretical frequencies respectively

The sampling distribution of this χ^2 -statistic was first obtained by *Helmert in 1875* and later on it was derived independently by *prof. K. Pearson in 1900*. The d.f. associated with χ^2 -distribution is the no. of cells or classes whose hypothetical frequencies can be determined independently. If the total no. of classes be n and k linear restrictions are imposed on them, the d.f. for χ^2 is taken as $(n - k)$. In short, the test is carried out as follows.

First we calculate the expected frequencies on the basis of the null hypothesis (H_0) and then compute the χ^2 statistic as—
 $\chi^2 = \Sigma(O - E)^2 / E$. If $\chi^2 \geq \chi^2_{0.05}(n - k)$, we reject the hyp. at 5 percent,

but if $\chi^2 < \chi^2_{.05} (n-k)$, there is no evidence against the hyp. at 5% level of significance.

Assumptions:—The above test is based upon the following assumptions or conditions—

- (1) The sample is large *i.e.* $\Sigma O \geq 50$.
- (2) The observations are independent.
- (3) The constraints if any, are always linear.

(4) None of the classes contains expected frequency < 5 . *Some statisticians take this no. as 10 also.*

In case the expected frequency of any cell is less than 5, regrouping is carried on with the neighbouring classes to make the frequencies > 5 .

Properties of χ^2 :

- (1) If $\chi^2 = 0$, there is a perfect agreement between the theory and practice. But as the value of χ^2 deviates from zero, it indicates a departure from this agreement. The greater the departure the larger is the value of χ^2 .
- (2) For large d. f. ($v \geq 30$), χ^2 —tends to normality.
- (3) **Additive Property :** If $\chi^2_1, \chi^2_2, \dots, \chi^2_k$ are all distributed independently like χ^2 -distribution with v_1, v_2, \dots, v_k , respective d.f. then the sum : $\chi^2_1 + \chi^2_2 + \dots + \chi^2_k$ will also be distributed like χ^2 with d. f. $v = v_1 + v_2 + \dots + v_k$.
- (4) **Partitioning Property :** Any χ^2 -variate with v d. f. can be split up into its independent component χ^2 -variables such that

$$\chi^2(v) = \chi^2(v_1) + \chi^2(v_2) + \dots + \chi^2(v_k), \text{ where } v = v_1 + v_2 + \dots + v_k.$$

Uses of χ^2 -test :

- (1) It is used to test whether a given sample has been taken from a population of specified variance.
- (2) It is used to test the homogeneity of several variances through the *Bartlett-test*.
- (3) It is used as a *test of goodness of fit* for testing whether a given sample has been taken from a specified population or distribution.
- (4) It is used to test the association of attributes.
- (5) It is used to test the homogeneity of several observed correlation coefficients.
- (6) It is extensively used in the analysis of *Genetical-Experiments* especially to test the genetical hypotheses and to detect the linkage.

5.6.1. Comparison of sample-variance with the population-variance :

If we want to test the hyp.—*that the sample has been taken from a population of specified variance (σ^2)*, we compute the statistic $\chi^2 = \Sigma(x - \bar{x})^2 / \sigma^2$.

It follows a χ^2 -distribution with $\nu = (n-1)$ d.f., where 'x' stands for the sample-observations, ' \bar{x} ' for the sample-mean and ' Σ ' for the summation over all 'n' values of the sample. If the population-mean ' μ ' is known in a problem, the proper statistic to be computed there for the purpose is $\chi^2 = \Sigma(x - \mu)^2 / \sigma^2$ which follows a χ^2 -distribution with $\nu = n$ d.f.

If $\chi^2 \geq \chi^2_{.05}(\nu)$ we reject the hyp. at 5% level,

but if $\chi^2 < \chi^2_{.05}(\nu)$, there is no evidence against the hyp. at 5% level of significance.

Exp. (13) For a sample of 10, the sum of squares of the deviations from the sample mean is 165.65. Test whether the sample has been taken from a normal population with s.d. ' $\sigma = 3$ ' ?

Sol. (H_0) : *The sample has been taken from the population of specified variance : $\sigma^2 = 9$.*

Here we compute the statistic

$$\chi^2 = \Sigma(x - \bar{x})^2 / \sigma^2 = 165.65 / 9.$$

$$\text{or } \chi^2 = 18.406 \text{ and } \chi^2_{.05}(9) = 16.919.$$

$\therefore \chi^2 > \chi^2_{.05}(9)$ leading to the rejection of H_0 at 5% level.

Conclusion : The sample has not been taken from the population of specified variance.

5.6.2 Testing the homogeneity of several variances :

If we want to test the hyp.—*that all the k-samples have been taken from the populations of equal variances i.e.*

$\sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2$, we compute Bartlett's-statistic as

$$\chi^2 = [n \log_e S_a^2 - \Sigma \nu \log_e S^2] / \left[1 + \frac{1}{3(k-1)} \left\{ \Sigma \left(\frac{1}{\nu} \right) - \frac{1}{n} \right\} \right].$$

It follows a χ^2 -distribution with $(k-1)$ d.f;

where k = no. of samples under consideration,

$$n = \nu_1 + \nu_2 + \dots + \nu_k, \nu_1 = n_1 - 1, \nu_2 = n_2 - 1, \dots, \nu_k = n_k - 1,$$

n_1 = size of the 1st. sample, n_2 = size of the 2nd-sample,

$$S_a^2 = (\nu_1 S_1^2 + \nu_2 S_2^2 + \dots + \nu_k S_k^2) / n.$$

$$S_1^2 = \Sigma(x_1 - \bar{x}_1)^2 / \nu_1; S_2^2 = \Sigma(x_2 - \bar{x}_2)^2 / \nu_2, \dots S_k^2 = \Sigma(x_k - \bar{x}_k)^2 / \nu_k$$

$$\Sigma \nu \log_e S^2 = \nu_1 \log_e S_1^2 + \nu_2 \log_e S_2^2 + \dots + \nu_k \log_e S_k^2,$$

$$\text{and } \Sigma \left(\frac{1}{\nu} \right) = \frac{1}{\nu_1} + \frac{1}{\nu_2} + \dots + \frac{1}{\nu_k}.$$

If $\chi^2 \geq \chi^2_{.05}(k-1)$, we reject the hyp. at 5% level,

but if $\chi^2 < \chi^2_{.05}(k-1)$, there is no evidence against the hyp. at 5% level of significance.

5.6.3 Testing the independence of attributes in contingency tables :

If the data is classified into m -rows and n -columns representing the m -classes according to one attribute, say A , and n -classes according to the other attribute, say B , thus in all into $m \times n$ classes, then such a table is known a $m \times n$ contingency-table as given below.

$B \backslash A$	B_1	...	B_j	...	B_n	Totals
A_1						R_1
\vdots			\vdots
A_i			E_{ij}			R_i
\vdots			\vdots
A_m						R_m
Totals	C_1	...	C_j	...	C_n	N

Now, if we want to test the hyp.—that these ' mn ' classes are independent, we calculate first on the basis of the hyp. of independence the expected frequencies for all the cells and then compute the desired statistic- χ^2 . The expected frequencies are calculated by the formula—

$E_{ij} = (R_i \times C_j) / N$, where R_i stands for the total of i -th row, C_j for the total of j -th-column, N for the grand total i.e. $N = \sum R_i (i=1, 2, \dots, m) = \sum C_j (j=1, 2, \dots, n)$, and E_{ij} for the expected frequency which falls into the cell of i -th-row and j -th-column. The desired statistic- χ^2 is computed as

$\chi^2 = \sum (O - E)^2 / E$, where O , E stand for observed, expected frequencies respectively.

If $\chi^2 \geq \chi^2_{.05}(v)$, we reject the hyp. at 5% level; $v = (m-1)(n-1)$, but if $\chi^2 < \chi^2_{.05}(v)$, there is no evidence against the hyp. at 5% level of significance.

Note :—In a contingency-table, the expected frequencies for the cells of the last row and last column should always be obtained by subtraction. The d.f. associated with a χ^2 -computed from a $m \times n$ contingency-table is the no. of cells whose expected frequencies can be calculated independently i.e. $v = (m-1)(n-1)$.

Exp. (14) Classification of fields as irrigated and unirrigated in a crop-cutting survey on wheat in four districts gave the following results—

District	: A	B	C	D	Totals
Irrigated	: 44	36	12	30	122
Unirrigated	: 130	167	98	143	538

Totals 174 203 110 173 | 660. Are these districts homogeneous in regard to proportion of irrigated fields of wheat ?

(M. Sc. Ag. Agra, 1956)

Sol. As the process of testing the homogeneity is the same as that of testing the independence, so the present problem may be treated as the problem of a 2×4 contingency-table.

Ho : *The districts are homogeneous in regard to the proportion of irrigated fields of wheat i.e. the eight classes are independent.*

The computations for the expected frequencies and the desired χ^2 are shown in the following table.

Class No.	Obs. freq. O	Expected freq. E	O-E	(O-E) ²	$\frac{(O-E)^2}{E}$
1	44	$(174 \times 122)/660 = 33$	11	121	3.667
2	36	$(203 \times 122)/660 = 37$	- 1	1	0.027
3	12	$(110 \times 122)/660 = 20$	- 8	64	3.200
4	30	$122 - (33 + 37 + 20) = 32$	- 2	4	0.125
5	130	$174 - 33 = 141$	-11	121	0.858
6	167	$203 - 37 = 166$	1	1	0.006
7	98	$110 - 20 = 90$	8	64	0.711
8	143	$173 - 32 = 141$	2	4	0.028
Totals	660 = N	N = 660	-	-	$\frac{8.622}{E} = \Sigma \frac{(O-E)^2}{E}$

Now we see that $\chi^2 = 8.622$ and $\chi^2_{.05} (2-1) (4-1) = 7.815$.

So $\chi^2 > \chi^2_{.05} (3)$ leading to the rejection of H_0 at 5% level.

Conclusion : The districts are not homogeneous in regard to the proportion of irrigated fields of wheat.

Exp. (15) : A new vaccine was tried on a certain no. of animals in a herd of cattle which was exposed to a certain disease. The nos. affected and not-affected in the categories vaccinated and not-vaccinated were as follows—

	Affected	Not-affected	Totals
Vaccinated :	25	75	100
Not-vaccinated :	15	85	100
Totals	40	160	200

Is the vaccine useful ?

(M. Sc. Ag. Agra, 1958)

Sol. (H_0) : *The vaccine is not useful i.e. the two characters viz. vaccination and effect are independent.*

The computations for the expected frequencies and the desired χ^2 are shown in the following table.

Class	O	E	O-E	(O-E) ²	$\frac{(O-E)^2}{E}$
1. Vac. affected	25	$(40 \times 100)/200 = 20$	5	25	1.2500
2. Vac. not. aff.	75	$100 - 20 = 80$	-5	25	0.3125
3. Not. vac. aff.	15	$40 - 20 = 20$	-5	25	1.2500
4. Not. vac. Not. aff.	85	$160 - 80 = 80$	5	25	0.3125
Totals	200 = N	N = 200	—	—	$\frac{3.1250}{\Sigma(O-E)^2/E}$

Now we see that $\chi^2 = 3.125$ and $\chi^2_{.05}(2-1) (2-1) = 3.841$.

So $\chi^2 < \chi^2_{.05}(1)$ leading to the acceptance of H_0 at 5% level.

Conclusion : The vaccine is not useful in controlling the disease to the animals.

5.6.3.1 Alternative method for computing χ^2 for a 2×2 contingency-table :

If we have a problem of a 2×2 contingency-table of the type $\frac{a|b}{c|d}$

as given below. Then to test the hyp.—

that the two characters are independent,
we can compute the statistic

$$\chi^2 = \frac{(ad - bc)^2 \times N}{R_1 R_2 C_1 C_2} \quad (\text{provided none of the class frequencies} < 5),$$

$$\text{or } \chi^2 = \left(\left| ad - bc \right| - \frac{N}{2} \right)^2 \times N$$

$$R_1 R_2 C_1 C_2$$

(provided any of the class-frequencies < 5).

B \ A	B ₁	B ₂	Totals
A ₁	a	b	R ₁
A ₂	c	d	R ₂
Totals	C ₁	C ₂	N

If $\chi^2 \geq \chi^2_{.05}(1)$, we reject the hyp. at 5% level,
but if $\chi^2 < \chi^2_{.05}(1)$, there is no evidence against the hyp. at 5% level of significance.

Exp. (16) : In an orchard of 1000 trees, a data was taken of the no. of shaded and unshaded trees, and in each of the classes the proportion of high to low yielding trees was noted. The results were recorded as in the following table.

	Shaded	Un-shaded	
High-yielders :	350	205	
Low-yielders :	250	195	Do these figures show that the shade has any effect on the yield of the trees ?

(M. Sc. Ag. Agra, 1961)

Sol. (H₀) : *The shade has no effect on the yield of the trees.*

The present problem can be arranged into a 2×2 contingency table and the desired χ^2 -statistic can be computed directly as shown below.

$$\chi^2 = \frac{(ad-bc)^2 \times N}{R_1 R_2 C_1 C_2}$$

$$= \frac{(50 \times 195 - 205 \times 250)^2 \times 1000}{555 \times 445 \times 600 \times 400}$$

$$= \frac{144500}{29639}$$

or $\chi^2 = 4.87$ and $\chi^2_{.05}(1) = 3.841$.
So $\chi^2 > \chi^2_{.05}(1)$ leading to the rejection of H_0 at 5% level.

Shade Yield \	Shaded	Un-Shaded	Totals
High	350 =a	205 =b	555 =R ₁
Low	250 =c	195 =d	445 =R ₂
Totals	600 =C ₁	400 =C ₂	1000 =N

Conclusion : The shade has an effect on the yield of the trees i.e. the two characters viz. *shade* and *yield* are not independent.

Exp. (17) In an experiment on the immunization of goats from Anthrax, the following results were obtained.

	Dead	Survived	
Inoculated :	2	10	
Not-inoculated :	6	6	Derive your inference on the efficiency of the vaccine ?

Sol. (H₀) : *The vaccination has no effect on the survival of goats from Anthrax.*

The data can be arranged into a 2×2 contingency-table and the desired χ^2 -statistic can be computed directly with *Yate's correction*, as shown below.

$$\begin{aligned}\chi^2 &= \left(\left| ad - bc \right| - \frac{N}{2} \right)^2 \times N \\ &\quad \frac{R_1 R_2 C_1 C_2}{(| 2 \times 6 - 10 \times 6 | - 12)^2 \times 24} \\ &= \frac{(48 - 12)^2 \times 24}{12 \times 12 \times 8 \times 16} \\ &= \frac{36 \times 36 \times 24}{12 \times 12 \times 8 \times 16} \\ &= \frac{27}{16}\end{aligned}$$

Sur. Vac.	Dead	Survived	Totals
Inocu.	2 =a	10 =b	12 =R ₁
Not-inocu.	6 =c	6 =d	12 =R ₂
Totals	8 =C ₁	16 =C ₂	24 =N

or $\chi^2 = 1.6875$ and $\chi^2_{.05}(1) = 3.841$.

So $\chi^2 < \chi^2_{.05}(1)$ leading to the acceptance of H_0 at 5% level.

Conclusion : The vaccination has no effect on the survival of goats from Anthrax *i.e.* the two characters viz. *vaccination* and *survival* are independent.

5.6.4 Applications of χ^2 in genetical-experiments :

5.6.4. (a) Testing the significance of ratio in a single-factor segregation : Sometimes in a genetical problem of single factor segregation, we want to test the significance of deviation of an observed segregation from a theoretical one *i.e.* the hyp. (H_0)-*that the classes or genes of a factor segregate in the given ratio.* For the purpose, first we calculate on the basis of the hyp. the expected frequencies of the classes and then compute the statistic

$\chi^2 = \sum (O - E)^2 / E$, where the symbols have their usual meanings

If $\chi^2 \geq \chi^2_{.05}(v)$, we reject the hyp. at 5% level,

but if $\chi^2 < \chi^2_{.05}(v)$, there is no evidence against the hyp. at 5% level of *significance*.

Exp. (18) : In a single factor F_2 -cross ($Aa \times Aa$), the observed class frequencies for the two different classes 'A' and 'a' are 312 and 88 respectively. Test, whether the data is in agreement with the Mendelian ratio 3 : 1 ?

Sol. (H_0) : The data is in agreement with the ratio 3 : 1.

Tests of Significance

The computations for the expected frequencies and the desired χ^2 are shown in the following table.

Class	O	E	O-E	(O-E) ²	$\frac{(O-E)^2}{E}$
A	312	$(3 \times 400)/4 = 300$	12	144	0.48
a	88	$400 - 300 = 100$	-12	144	1.44
Totals	400 =N	N = 400	—	—	1.92 = $\Sigma(O-E)^2/E$

Now we have $\chi^2 = 1.92$ and $\chi^2_{.05}(1) = 3.841$.

So $\chi^2 < \chi^2_{.05}(1)$ leading to the acceptance of H_0 at 5% level.

Conclusion : The observed data are in agreement with the ratio 3 : 1.

Exp. (19) In a cross between ivory and red snapdragons, an experimenter obtained the following results in the F_2 -generation—

Phenotype	Red	Pink	Ivory
No. of plants	22	52	23

Test whether these figures show that the segregation occurs in a Mendelian ratio of 1 : 2 : 1 ?

Sol: (H_0) : *The segregation occurs in the ratio 1 : 2 : 1 in the observed data.*

The computations for the expected frequencies and the desired χ^2 are shown in the following table.

Class	O	E	O-E	(O-E) ²	$(O-E)^2/E$
Red	22	$(1 \times 97)/4 = 24.25$	-2.25	5.0625	0.2087
Pink	52	$(2 \times 97)/4 = 48.50$	3.5	12.2500	0.2525
Ivory	23	$97 - (24.25 + 48.50) = 24.25$	1.25	1.5625	0.0644
Totals	97 =N	N = 97	—	—	0.5256 = $\Sigma(O-E)^2/E$

Now we have $\chi^2 = 0.526$ and $\chi^2_{.05}(2) = 5.991$.

So $\chi^2 < \chi^2_{.05}(2)$ leading to the acceptance of H_0 at 5% level.

Conclusion : The observed data are in agreement with the ratio 1 : 2 : 1.

5. 6. 4. (b) Testing the homogeneity of several families in a single factor segregation :

Sometimes in a genetical problem of single factor segregation,

the data are available for a no. of families 'say k' each with two genes or classes "say A, a" segregating in the given ratio 'say $m_1 : m_2$ '.

Here we may be interested in knowing the *consistency of the families* or groups. In such cases, the value of χ^2 is calculated separately for each family and all these values are added up to get the *total chisquare* 'say χ^2_T ' with k. d. f. This χ^2_T is further partitioned into the following two independent χ^2 -components.

Family	Genes		Totals
	A	a	
1	A_1	a_1	R_1
2	A_2	a_2	R_2
...
k	A_k	a_k	R_k
Totals	C_1	C_2	N

(1) χ^2_D : due to deviations from the theoretical ratio.

(2) χ^2_H : due to heterogeneity between the families.

Family	Computed χ^2	D.F.	$\chi^2_{.05}$
1	$\chi^2_1 = \frac{(m_2 A_1 - m_1 a_1)^2}{m_1 m_2 R_1} = \dots$	1	3.841
2	$\chi^2_2 = \frac{(m_2 A_2 - m_1 a_2)^2}{m_1 m_2 R_2} = \dots$	1	"
...
k	$\chi^2_k = \frac{(m_2 A_k - m_1 a_k)^2}{m_1 m_2 R_k} = \dots$	1	"
Totals	$\chi^2_T = \chi^2_1 + \chi^2_2 + \dots + \chi^2_k = \dots$	k	...
Due to dev.	$\chi^2_D = \frac{(m_2 C_1 - m_1 C_2)^2}{m_1 m_2 N} = \dots$	1	3.841
Due to hetero.	$\chi^2_H = \chi^2_T - \chi^2_D = \dots$	k-1	...

The first component ' χ^2_D ' with 1 d. f. is obtained from the totals of the two genes over all the families while the second component ' χ^2_H ' with (k-1) d. f. is obtained by subtracting χ^2_D from χ^2_T . These two component-chisquares viz. χ^2_D and χ^2_H are used to test the following (i) and (ii) hypotheses respectively.

H_0 : (i) The data are in agreement with the given ratio ' $m_1 : m_2$ '.

(ii) The families are homogeneous in regard to the segregation ratio ' $m_1 : m_2$ '.

Finally, we arrive at the following results—

(1) If $\chi^2_D \geq \chi^2_{.05}(1)$, we reject the hyp. (i) at 5% level i.e. the families do not segregate in the given ratio,

or the data on an average do not show an agreement with the given ratio ' $m_1 : m_2$ '.

But if $\chi^2_D < \chi^2_{.05}(1)$, there is no evidence against the hyp. (i) at 5% level i.e. the data show an agreement with the given ratio ' $m_1 : m_2$ '.

(2) If $\chi^2_H \geq \chi^2_{.05}(k-1)$, we reject the hyp. (ii) at 5% level i.e. the families are not homogeneous in regard to the segregation ratio ' $m_1 : m_2$ '.

But if $\chi^2_H < \chi^2_{.05}(k-1)$, there is no evidence against the hyp. (ii) at 5% level i.e. the families are homogeneous in regard to the segregation ratio ' $m_1 : m_2$ '.

Exp (20) : For the data given in the table, test whether

(i) the genes 'A' and 'a' segregate in the ratio 3 : 1, and

(ii) the families are homogeneous in the segregation of (A,a) in the ratio 3 : 1 ?

Sol. (H_0) : (i) The genes (A, a) segregate in the ratio 3 : 1.

(ii) The families are homogeneous in regard to the segregation ratio 3 : 1.

Genes Family	A		Totals
	A	a	
1	55	20	75
2	89	25	114
3	78	23	101
4	61	19	80
Totals	283	87	370

The computations for the desired statistics viz, χ^2_T , χ^2_D and χ^2_H are shown in the following table.

Family	Computed χ^2	D.F.	$\chi^2_{.05}$
1	$\chi^2_1 = \frac{(1 \times 55 - 3 \times 20)^2}{3 \times 1 \times 75} = 0.1111$	1	3.841
2	$\chi^2_2 = \frac{(1 \times 89 - 3 \times 25)^2}{3 \times 1 \times 114} = 0.5731$	1	"
3	$\chi^2_3 = \frac{(1 \times 78 - 3 \times 23)^2}{3 \times 1 \times 101} = 0.2673$	1	"
4	$\chi^2_4 = \frac{(1 \times 61 - 3 \times 19)^2}{3 \times 1 \times 80} = 0.0667$	1	"
Totals	$\chi^2_T = \chi^2_1 + \dots + \chi^2_4 = 0.0182$	4	9.488
Due to ev.	$= \frac{(1 \times 283 - 3 \times 87)^2}{3 \times 1 \times 370} = 0.0436$	1	3.841
Due to hetero.	$\chi^2_H = \chi^2_T - \chi^2_D = 0.9646$	3	7.815

Now we see that—

(i) $\chi^2_D < \chi^2_{.05}$ (1) leading to the acceptance of hyp. (i) at 5% level.

(ii) $\chi^2_H < \chi^2_{.05}$ (3) leading to the acceptance of hyp. (ii) at 5% level.

Conclusion : The genes (A,a) segregate in the given ratio 3 : 1 and the families are homogeneous in regard to this segregation ratio.

Exp. (21) For the figures given in the table, test whether—

(a) : (i) the genes 'A' and 'a' segregate in the ratio 9 : 7, and

(ii) the families are homogeneous in the segregation of (A, a) in the ratio 9 : 7 ?

(b) : Also test the homogeneity of five families in the ratio whatever they indicate ?

Family	Genes		Totals
	A	a	
1	70	55	125
2	35	26	61
3	56	29	85
4	34	32	66
5	39	24	63
Totals	234	166	400

Sol. (a). (H_0) : (i) The genes (A,a) segregate in the ratio 9 : 7.

(ii) The families are homogeneous in regard to the segregation ratio 9 : 7.

The computations for the desired statistics viz. χ^2_T , χ^2_D and χ^2_H are shown in the following table.

Family	Computed χ^2	D. F.	$\chi^2_{.05}$
1	$\chi^2_1 = \frac{(7 \times 70 - 9 \times 55)^2}{9 \times 7 \times 125} = 0.0032$	1	3.841
2	$\chi^2_2 = \frac{(7 \times 35 - 9 \times 26)^2}{9 \times 7 \times 61} = 0.0315$	1	"
3	$\chi^2_3 = \frac{(7 \times 56 - 9 \times 29)^2}{9 \times 7 \times 85} = 3.2047$	1	"
4	$\chi^2_4 = \frac{(7 \times 34 - 9 \times 32)^2}{9 \times 7 \times 66} = 0.6013$	1	"
5	$\chi^2_5 = \frac{(7 \times 39 - 9 \times 24)^2}{9 \times 7 \times 63} = 0.8186$	1	"
Totals	$\chi^2_T = \chi^2_1 + \dots + \chi^2_5 = 4.659$	5	11.070
Due to dev.	$\chi^2_D = \frac{(7 \times 234 - 9 \times 166)^2}{9 \times 7 \times 400} = 0.8229$	1	3.841
Due to hetero.	$\chi^2_H = \chi^2_T - \chi^2_D = 3.8364$	4	9.488

Now we see that—

- (i) $\chi^2_D < \chi^2_{.05} (1)$ leading to the acceptance of hyp. (i) at 5% level;
 (ii) $\chi^2_H < \chi^2_{.05} (4)$ leading to the acceptance of hyp. (ii) at 5% level.

Conclusion : The genes (A,a) segregate in the given ratio 9 : 7 and the families are homogeneous in regard to this segregation ratio.

(b). H_0 : *The families are homogeneous in regard to the segregation ratio whatever the data indicate i.e. 234 : 166.*

The testing of *homogeneity between the families* in the ratio whatever they indicate in the observed data is the same as that of testing the *independence of two characters*. Hence, the question will be answered in the same way as that of testing the independence in a $n \times 2$ contingency table. The computations for the expected frequencies and the desired χ^2 are shown in the following table.

Class	O	E	O-E	(O-E) ²	$\frac{(O-E)^2}{E}$
Family 1.A	70	$(234 \times 125)/400 = 73$	-3	9	0.1233
„ 2.A	35	$(234 \times 61)/400 = 36$	-1	1	0.1278
„ 3.A	56	$(234 \times 85)/400 = 50$	6	36	0.7200
„ 4.A	34	$(234 \times 66)/400 = 39$	-5	25	0.6410
„ 5.A	39	$234 - (73 + 36 + 50 + 39) = 36$	3	9	0.2590
„ 1.a	55	$125 - 73 = 52$	3	9	0.173
„ 2.a	26	$61 - 36 = 25$	1	1	0.0400
„ 3.a	2	$85 - 50 = 35$	-6	36	1.0286
„ 4.a	32	$66 - 39 = 27$	5	25	0.9259
„ 5.a	34	$63 - 36 = 27$	-3	9	0.333
Totals	400 = N	N=400	—	—	$4.2630 = \Sigma(O-E)^2/E$

Now we have $\chi^2 = 4.263$ and $\chi^2_{.05} (4) = 9.488$.

So $\chi^2 < \chi^2_{.05} (4)$ leading to the acceptance of hyp. at 5% level.

Conclusion : The families are homogeneous in regard to the segregation ratio 234 : 166.

5 6. 4. (c) Testing the significance of ratio in two factor-segregation :

Sometimes in a genetical problem of two factor segregation, the data are available for two characters or factors 'say A, B' each with two classes or genes 'say A, a and B, b respectively' segregating simultaneously in the given ratios 'say $m_1 : 1$ and $m_2 : 1$ respectively. Here we may be interested in knowing *whether the two factors are segregated in the given ratios independently, or linked*. Thus the problem may be looked as testing the hypotheses—

- H_0 (i) *The genes (A, a) are segregated in the given ratio ' $m_1 : 1$ '.*
 (ii) *The genes (B, b) are segregated in the given ratio ' $m_2 : 1$ '.*
 (iii) *Two factors are independently segregated in the given ratios.*

Here on the basis of the hyp. of independence of the two characters and their segregation according to the given ratios, we obtain the four classes viz. AB, Ab, aB, and ab. The frequencies of these four classes are supposed to be in the ratios $m_1 m_2 : m_1 : m_2 : 1$ respectively. Thus if the *observed frequencies* of the above mentioned four classes are (in the order) a_1, a_2, a_3 and a_4 with $N = (a_1 + a_2 + a_3 + a_4)$, then their corresponding *expected frequencies* may be obtained as shown below.

$$E(a_1) = \frac{m_1 m_2 N}{(m_1 + 1)(m_2 + 1)}, \quad E(a_2) = \frac{m_1 N}{(m_1 + 1)(m_2 + 1)},$$

$E(a_3) = \frac{m_2 N}{(m_1 + 1)(m_2 + 1)}$, and $E(a_4) = N - [E(a_1) + E(a_2) + E(a_3)]$. Now we compute the desired statistic $\chi^2 = \sum (O - E)^2 / E$ by the usual method, which follows a χ^2 -distribution with 3 d. f.

If $\chi^2 < \chi^2_{.05}(3)$, there is no evidence against the hyp. (H_0) at 5% level. It concludes that the two characters are independent and also they are segregated in the given ratios.

It should be noted that in such a situation the answer is complete with this stage. But in the contrary case, the procedure is still carried on for further detecting the cause of its being significant, as stated below.

If $\chi^2 \geq \chi^2_{.05}(3)$, we reject the hyp. at 5% level of significance.

In this situation, the significant value of χ^2 may be supposed to be so owing to some discrepancy in the hyp. (H_0), which may be due to any one or more of the following three reasons.

- (1) The genes (A, a) might have not segregated in the given ratio.
- (2) The genes (B, b) might have not segregated in the given ratio.
- (3) The two characters might have not segregated in the given ratios independently but might have given some evidence of linkage.

Thus in order to detect the above mentioned causes of discrepancy, we partition the above calculated χ^2 into three independent component-chisquares each with 1 d. f. as shown in the following table.

Source due to	Computed χ^2	D. F.	
A-Character	$\chi^2_A = \frac{[a_1 + a_2 - m_1(a_3 + a_4)]^2}{m_1 N} = \dots$	1	3.841
B-Character	$\chi^2_B = \frac{[a_1 + a_3 - m_2(a_2 + a_4)]^2}{m_2 N} = \dots$	1	„
L-Linkage	$\chi^2_L = \chi^2 - (\chi^2_A + \chi^2_B) = \dots$ or $\frac{[a_1 + m_1 m_2 a_4 - (m_2 a_2 + m_1 a_3)]^2}{m_1 m_2 N}$	1	„
Totals	$\chi^2 = \chi^2_A + \chi^2_B + \chi^2_L = \dots$	3	7.815

These three component-chisquares viz. χ^2_A , χ^2_B and χ^2_L are used to test the hyp. (i), (ii) and (iii) respectively.

Finally, we arrive at the following results—

(1) If $\chi^2_A \geq \chi^2_{.05}(1)$, we reject the hyp. (i) at 5% level i.e. the genes (A, a) are not segregated in the given ratio.

But if $\chi^2_A < \chi^2_{.05}(1)$, there is no evidence against the hyp. (i) at 5% level i.e. the genes (A, a) are segregated in the given ratio.

(2) If $\chi^2_B \geq \chi^2_{.05}(1)$, we reject the hyp. (ii) at 5% level i.e. the genes (B, b) are not segregated in the given ratio.

But if $\chi^2_B < \chi^2_{.05}(1)$, there is no evidence against the hyp. (ii) at 5% level i.e. the genes (B, b) are segregated in the given ratio.

(3) If $\chi^2_L \geq \chi^2_{.05}(1)$, we reject the hyp. (iii) at 5% level i.e. the two characters are not segregated in the given ratios independently but have an evidence of linkage.

But if $\chi^2_L < \chi^2_{.05}(1)$, there is no evidence against the hyp. (iii) at 5% level i.e. the two characters are segregated in the given ratios independently.

Exp. (22) In a cross involving two Mendelian factors, the following results were obtained in the F_2 segregation. Are these observations in accordance with the hypothesis that the two factors segregate independently and that the four classes of offsprings are equally viable ?

(M. Sc. Ag. Agra, 1951)

Flat leaves		Crimpled leaves	
Normal eye	Primrose eye	Normal eye	Primrose eye
328	122	77	33

Sol. (H_0): The two factors viz. the shape of leaves and the colour of eye are segregated independently in the ratio 3 : 1 each.

Let the two factors be A and B. Also, let 'A' stand for flat leaves, 'a' for crimped leaves, 'B' for normal eye and 'b' for primrose eye. Now on the basis of the hyp. assumed, if (A, a) are segregated in the ratio 3 : 1, (B, b) are also segregated in the ratio 3 : 1, and the two characters are segregated in the given ratios independently, then the frequencies of the four classes AB, Ab, aB, ab will occur in the ratios 9 : 3 : 3 : 1 respectively. The expected frequencies and the desired χ^2 are computed in the following table.

Class	O	E	O-E	(O-E) ²	(O-E) ² /E
AB	328	$\frac{9}{16} \times 560 = 315$	13	169	0.5365
Ab	122	$\frac{3}{16} \times 560 = 105$	17	289	2.7524
aB	77	$\frac{3}{16} \times 560 = 105$	-28	784	7.4667
ab	33	$560 - (315 + 105 + 105) = 35$	-2	4	0.1143
Totals	$\frac{560}{=N}$	N = 560	—	—	$\frac{10.8699}{= \Sigma(O-E)^2/E}$

Now we have $\chi^2 = 10.8699$ and $\chi^2_{.05}(3) = 7.815$.

So $\chi^2 > \chi^2_{.05}(3)$ leading to the rejection of hyp. at 5% level of significance. It shows some discrepancy in the hyp. assumed. In order to detect the cause of this discrepancy, we partition the above calculated χ^2 into three independent component-chisquares

viz. χ^2_A , χ^2_B and χ^2_L as shown in the following table.

Source due to	Computed χ^2	D.F.	$\chi^2_{.05}$
A-Shape of leaves	$\chi^2_A = \frac{[328 + 122 - 3(77 + 33)]^2}{3 \times 560} = 8.5714$	1	3.841
B-Colour of eye	$\chi^2_B = \frac{[328 + 77 - 3(122 + 33)]^2}{3 \times 560} = 2.1429$	1	„
L-Linkage	$\chi^2_L = 10.8699 - (8.5714 + 2.1429) = 0.1556$	1	„
Totals	$\chi^2 = 10.8699$	3	7.815

Now we see that—

- (i) $\chi^2_A > \chi^2_{.05}$ (1) leading to the rejection of hyp. at 5% level,
- (ii) $\chi^2_B < \chi^2_{.05}$ (1) „ „ „ acceptance of „ „ „ , and
- (iii) $\chi^2_L < \chi^2_{.05}$ (1) „ „ „ „ „ „ „ „ „ 5% level.

Conclusions :

- (1) The character '*shape of leaves*' is not segregated in the ratio 3 : 1.
- (2) The character '*colour of eye*' is segregated in the ratio 3:1.
- (3) The test provides no evidence of linkage i.e. the two characters are segregated independently.

EXERCISE V

Q. 1. A sample of 100 plants was found to have a mean height of 73.65 cms. Could it be reasonably regarded as a simple random sample from a normal population whose mean is 75 cms. and s.d. is 3.0 cms. ?

Q.2 Suppose that 60 senior students in a college A and 80 senior students in another college B had mean statures of 69 and 67.5 inches respectively. If the s. d. for statures of all the senior students is 2.40 inches, is the difference between the mean statures of the two groups is significant at 1 percent level of significance ? Given SNV 'Z' = 2.58.

Q.3 A potential buyer of light bulbs bought 50 bulbs of each of two brands A and B. Upon testing these bulbs he found that brand A had a mean life of 1285 hours with a s.d. of 40 hours whereas brand B had a mean life of 1320 hours with a s.d. of 47 hours. Can the buyer be quite certain that the two brands A and B do differ in quality ?

Q. 4 A sample of 400 men from south India has a mean height of 65.85 inches and a s.d. of 2.50 inches, while a sample

of 100 men from North India has a mean height of 66.20 inches with a s.d. of 2.52 inches. Do the data indicate that the North Indians are on the average taller than the South Indians ?

Q.5. A random sample of 200 villages was taken from Gorakhpur District and the average population per village was found to be 485 with a s.d. of 50. Another random sample of 200 villages from the same district gave an average population of 510 per village with a s.d. of 40. Is the difference between the averages of two samples statistically significant ? Give reasons.

Q.6 (a) A random sample of 1000 farms in a certain year in Punjab gives an average yield of wheat of 20 mds./acre with a s.d. of 1.92 mds./acre. Another random sample of 1000 farms in the same year in U. P. gives an average yield of wheat of 21 mds./acre with a s.d. of 2.24 mds./acre. Show that the average yields in the two provinces are the same ?

(b) Explain the terms *null hypothesis* and *5 percent level of significance* ?

Q.7 Find Student's '*t*' for the following sample of eight drawn from a universe with zero mean—

−4, −2, −2, 0, 2, 2, 3, 3.

Q.8 A certain drug administered to each of 12 patients resulted in the following increases of blood-pressures—
5, 3, 8, −1, 3, 0, 6, −2, −1, 5, 0, 4.

Can it be concluded that the stimulus (drug) will in general be accompanied by an increase in blood-pressure ?

Q.9. The yields of two types 'Type 17' and 'Type 51' of grains in pounds per acre in 6 replications are given below. What comment would you make on the differences in the mean yields ? You may assume that if there be 5 d.f. and $p=0.2$, t is 1.476.

Replications	1	2	3	4	5	6
Type 17	: 20.50,	24.60	23.06	29.98	30.37,	23.83
Type 51	: 24.86,	26.39	28.19	30.75	29.97	22.04.

Q.10. Determination of protein content of five varieties of wheat by a standard and newly developed rapid method gave the following results (in units of grams per 100 grams).

Varities:	A	B	C	D	E
Standard Method :	13.5	13.0	13.0	13.6	14.1
Rapid Method :	13.1	13.1	12.8	13.1	13.7.

Do the estimates obtained by two methods differ significantly ?

Q. 11. Test whether a small electric current affects the growth of maize-seedlings. Ten pairs of plants were grown in parallel

boxes and one member of each pair was treated by receiving a small electric current. The differences in height (in mm.) between the treated and untreated were as follows.

6.0, 1.3, 10.2, 23.9, 3.1, 6.8, -1.5, -14.7, -3.3, 11.1.

Q. 12. (a) The following figures give the percentage extension under a given load of two random samples of yarn, the first sample being taken before washing, the second after six washings—

Before washing : 12.3 13.7 10.4 11.4 14.9 12.6

After six washings : 15.7 10.3 12.6 14.5 12.6 13.8 11.9.

Is there any evidence that extensibility is affected by washing?

(b) In another experiment on the same type of yarn, six lengths of yarn were selected at random and each length was cut into two halves. One of the halves was tested for extension without washing, the other after six washings and the following percentage extensions were obtained—

Length :	1	2	3	4	5	6
Before washing :	13.9	12.5	11.0	11.8	10.8	14.6
After six washings :	14.7	12.1	13.2	13.6	11.5	15.4.

Is there any evidence that extensibility is affected by washing?

Q.13 Eight pots growing three wheat plants each were exposed to a high tension discharge while nine similar pots were enclosed in an earthen wire case. The number of tillers in each pot were as follows—

Caged :	17	26	18	25	27	28	26	23	17
Electrified :	16	16	22	16	21	18	15	20.	

See whether electrification exercises any real effect on the tillering by using *t* test of significance ? (I.C.A.R. 1956)

Q. 14. Two horses A and B were tested according to the time (in seconds) to run on a particular track with the results—

Horse A :	28	30	32	35	33	29	34
Horse B :	29	30	30	24	27	29.	

Test whether you can discriminate between the two horses ?

Q. 15 The following data represent the yields in bushels of Indian corn on ten sub-divisions of equal areas of two agricultural plots in which plot I was control plot treated the same as plot II, except for the amount of phosphorus applied as a fertilizer—

Plot I :	6.2	5.7	6.5	6.0	6.3	5.8	5.7	6.0	6.0	5.8
Plot II :	5.6	5.9	5.6	5.7	5.8	5.7	6.0	5.5	5.7	5.5.

Is there a significant difference between the yields on the two plots, using the difference between their means a criterion of judgement ?

Q. 16. Two samples of sizes 10 and 12 give the sum of squares of deviations from their respective means 60.3 and 61.2

as 180 and 132 respectively. Can they be regarded as drawn from the same normal population ?

Q. 17 Why there are different tests in Statistics for testing the significance of mean difference ?

Mitchell conducted a paired feeding experiment with pigs on the relative value of limestone and bone-meal for bone-development. The results are ash content in% of scapulus of pairs fed on limestone and bone-meal.

Pair :	1	2	3	4	5	6	7,	8
Limestone :	49.2	53.3	50.6	52.0	46.8	50.5	52.1	53.0
Bone-meal :	51.5	54.9	52.2	53.3	51.6	54.1	54.2	54.3.

Determine the significance of the difference between the means—

(1) by assuming that the values are paired, and

(2) by assuming that the values are not paired. (ICAR, 1956)

Q. 18 (a) Define Student's 't' and give its applications.

(b) The following yields (in pounds per plot) were obtained from five plots each of two varieties of wheat A and B—

Variety A : 12 10 12 13 13

Variety B : 8 9 11 10 11.

Test the significance of the difference between the varieties by means of both t and F tests, assuming (i) plots are paired, and (ii) they are independent, there being no correspondence between the plots of the two varieties ? (M. Sc. Ag. Agra, 1957)

Q.19 Ten determinations of a quantity subject of error are—

5.5, 6.7, 6.8, 6.4, 7.2, 6.9, 6.6, 5.8, 6.3, 6.5.

Judge whether the sample mean is compatible with a normal population with mean as 7.0 and s.d. unity. Show that your inference is reversed if population-variance be unknown ? (M.Sc.Ag. Agra, 1962)

Q.20 For two random samples of sizes 10 and 12, given that

$\bar{x}_1=22$, $\bar{x}_2=25$, $\Sigma (x_1-\bar{x}_1)^2=120$ and $\Sigma (x_2-\bar{x}_2)^2=314$. Test, whether the two samples are drawn from the same normal population.

Q. 21 Two halves of a field were sown with two varieties of wheat A & B. One hundred earheads were selected randomly from each half and no. of grains in each earhead counted. The data gave the following results—

Sample A : mean=30.5, s.e. of mean=0.5

Sample B : mean=32.1, s.e. of mean=0.6.

Test the significance of difference between the means and give your inference from results. What would you have concluded if the above values had been obtained from samples of 5 earheads each instead of 100 earheads ? (M. Sc. Ag. Agra, 1956)

Q.22 Lots of ten bees were fed two concentrations of syrup—20% and 65% at a feeder half a mile from the hive. Upon arrival at the hive their honey-sacs were removed and the concentration of fluid measured. In every case there was a decrease from feeder concentrations. The decrease were—

From 20% to syrup : 0.7, 0.5, 0.4, 0.7, 0.5, 0.4, 0.7, 0.4, 0.2, 0.5

From 65% to syrup : 1.7, 2.8, 2.2, 1.4, 1.3, 2.1, 0.8, 3.4, 1.9, 1.4.

Test whether the decrease in concentration during flight differs significantly with the two syrups ? (M. Sc. Ag. Agra, 1960)

Q. 23 (a) What is the difference between 't' and 'F' tests, and when are the two tests identical ? (M. Sc. Ag. M.U. 1967)

(b) Write a critical note on tests of significance and their uses in agricultural statistics ? (M. Sc. Ag. M.U. 1967)

Q. 24 Discuss the uses of χ^2 in Genetical-analysis ?

It is claimed that students from cities are more sociable than students from the villages. Test if this claim is valid for the data—

Social Non-social

Students from cities : 10 3

Students from villages : 2 15 (M.A. Patna, 1955)

Q. 25 The following data are observed for hybrids of *Datura*—

Flowers violet, fruits prickly ... 47

„ „ „ smooth ... 12

„ white „ prickly ... 21

„ „ „ smooth ... 3.

Using χ^2 -test, find the association between colour of flowers and character of fruit ? (M. Sc. Ag. Agra, 1956)

Q. 26 Two samples of polls of votes for two candidates A and B for a public office are taken, one from among residents of urban areas and the other from residents of rural areas—

Votes for area	A	B	Totals
Rural :	620	380	1000
Urban :	550	450	1000.

Examine whether the nature of the area is related to voting preference in this election ? (I.A.S. 1956)

Q. 27 The following data show the effect of vitamin B-deficiency on the sex ratio of the offsprings of rats. Is the effect significant ?

	Male	Female	Totals
Vitamin B deficient :	123	153	276
Vitamin B sufficient :	145	150	295.

Q. 28 Twelve inoculated experimental animals and the other

12 not-inoculated animals were exposed to the infection of a disease. The following frequencies of dead and surviving animals were noted in the two cases.

	<i>Dead</i>	<i>Survived</i>	<i>Totals</i>
Inoculated :	2	10	12
Not-inoculated :	8	4	12.

Can the inoculation be said to have an evidence of preventing disease?

Q. 29 In an experiment with immunization of cattle from tuberculosis, the following results were obtained—

	<i>Affected</i>	<i>Unaffected</i>
Inoculated :	12	2
Not-inoculated :	16	6.

Examine the effect of vaccine in controlling susceptibility to tuberculosis ? (I.A.S. 1948)

Q. 30 The following data relate to two types of twins—

	<i>Both right handed</i>	<i>One left and the other right handed</i>
Fraternal :	10	2
Identical :	5	5.

Examine whether identical twins are different from fraternal twins in having a lower proportion of cases of both members of the twins being right-handed ? (M. A. Patna, 1955)

Q. 31 In a public preference survey, the people interviewed were classified according to their opinion regarding intercaste-marriage and their age as follows—

	<i>Opinion</i>				<i>Age in years</i>			
	19-25	26-35	36-55	Over				
Unconditional support :	76	125	96	10				
Conditional support :	69	117	16	17				
Indifference :	14	27	35	4				
Conditional opposition :	60	168	210	46.				

Examine whether and in what way opinion regarding intercaste-marriage changes with age ? (M.A. Patna, 1953)

Q. 32 (a) In experiments on pea-breeding, Mendel obtained the following frequencies of seeds—

Round and yellow=315, Wrinkled and yellow=101

Round and green=108, Wrinkled and green=32.

Theory predicts that the frequencies should be in the proportions 9 : 3 : 3 : 1. Examine the correspondence between theory and experiment ? (M.Sc.Ag. Agra, 1952)

(b) Define χ^2 and write in brief the conditions or assumptions for the application of χ^2 -test ? What is meant by degrees of freedom ?

Q. 33 (a) From the following data for 4 segregating F_2 -families, test whether they agree with one another with a ratio 3 : 1 and also in the ratio whatever they indicate—

Family	Class	
	A	B
1	72	21
2	55	20
3	89	25
4	78	23.

(b) Define χ^2 test and write its applications ?

Q. 34 In a back-cross progeny ($pt/PT \times pt/pt$), the observed frequencies in the four classes are given below.

Class	PT	Pt	pT	pt	Is there
Observed freq.	190	38	35	204.	

any evidence for the existence of linkage in the coupling phase ?

Q. 35 In an F_2 -segregatin for two characters A and B, the number of plants observed in the classes AB, Ab, aB and ab were—

AB	Ab	aB	ab	Total
290	90	75	25	480.

Test whether each of the genes is segregating according to a 3 : 1 ratio and whether the two characters are segregating independently ? Do you think that there is any evidence of linkage ?

Q. 36 Genetic theory states that children having one parent of blood type M and the other of blood type N will always be one of the three types M, MN, N and that the proportions of three type will on average be as 1 : 2 : 1. A report states that out of 300 children having one M parent and one N parent, 30 percent were found to be type M, 45 percent type MN and remainder type N. Test the hypothesis by χ^2 test ?

Q. 37. In a F_2 segregating for two characters A and B the no. of plants observed in the classes AB, Ab, aB, and ab are as—

AB	Ab	aB	ab	Total
289	92	73	26	480.

Test whether each of the genes is segregating according to 3 : 1 ratio and whether two genes are segregating independently ? Do you think there is any evidence of linkage ?

(M.Sc.Ag. Agra 1958)

Q. 38 In the F_2 -generation of a cross between a two rowed variety of barley producing green seedlings (VVLgLg) with a six-rowed variety with light green seedlings (vv lglg), the following nos. of plants were obtained in the four phenotypic classes.

Phenotypic class :	VLg	Vlg	vLg	vlg
No. of plants :	281	59	60	58.

Calculate the

goodness of fit between the ratio observed and the ratio expected on the basis of independent inheritance of the two factor pairs ?

(M. Sc. Ag. AU, 1965)

Q. 3 In a herd of cattle composed of four different breeds of the total nos. of animals of each breed and of animals affected by a certain epidemic-disease in each breed are as follows—

Breed	A	B	C	D
Total no. of animals	200	250	300	250
No. of animals affected	11	27	38	24

Do the breeds differ significantly in their susceptibility to disease ? (M.Sc.Ag. AU, 1958)

Q. 40 It was observed that out of 463 smokers 55 were found to suffer from heart trouble, while from 337 non-smokers only 25 were found to be affected thus. Would you conclude from these data that smoking affects health ? (M Sc. Ag. AU, 1959)

Q. 41 From the following table showing the no. of plants having certain characters, test the hypothesis that the flower-colour is independent of flatness of leaves.

	Flat leaves	Curled leaves
White flowers :	99	36
Red flowers :	20	5.

(M.Sc.Ag. AU, 1957)

Q. 42 The following data occur in a memoir of Karl Pearson—

		<i>Eye colour in sons</i>	
		Not light	light
Eye colour { in fathers }	Not light	230	148
	light	151	471.

Test whether the colour of the son's eyes is associated with that of the father's. (IAS, 1942)

Q. 43 In an antimalarial campaign in a certain area of Baraut, quinine was administered to 500 persons out of a total population of 1000. The no. of fever cases is shown below —

Treatment :	Fever	No-Fever
Quinine :	200	300
No-Quinine :	150	350.

Test the usefulness of quinine in preventing malaria attack.

Q. 44 From the figures given below, test whether the intelligence of the sons is associated with that of the fathers.
Int. fathers with int. sons = 20, Int. fathers with dull sons = 30,
Dull fathers with int. sons = 40, Dull fathers with dull sons = 70.

Q. 45 The following table gives the no. of aircraft accidents that occurred during the various days of the week. Test whether the accidents are uniformly distributed over the week.

Days :	Sun.	Mon.	Tue.	Wed.	Thu.	Fri.	Sat.	Total
No. of accidents :	14	16	8	12	11	9	14	84.

Q. 46 Two hundred digits were chosen at random from a set of tables. The frequencies of the digits were as follows—

Digits : 0 1 2 3 4 5 6 7 8 9

Frequencies : 18 19 23 21 16 25 22 20 21 15. Use the χ^2 -test to assess the correctness of the hypothesis that the digits were distributed in equal nos. in the table from which these nos. were taken.

ANSWERS

(1) SNV ' z '=4.5 (2) SNV ' z '=3.67 (3) SNV ' z '=4.01 (4) SNV ' z '=1.25 (5) SNV ' z '=5.5 (6) [a]. SNV ' z '=10.75 (7) Student ' t '=0.266 (8) Student ' t '=2.688 (9) Paired ' t '=1.55 (10) Paired ' t '=2.99 (11) Paired ' t '=1.33 (12) [a]. Fisher ' t '=0.54, [b]. Paired ' t '=2.62 (13) Fisher ' t '=2.74 (14) Fisher ' t '=2.44 (15) Paired ' t '=2.63 (16) Fisher ' t '=0.53 (17) [i]. Paired ' t '=4.44 [ii]. Fisher ' t '=2.61 (18) [b]. (i) F=1.13, Paired ' t '=3.79, (ii) Fisher ' t '=5.5 (19) SNV ' z '=1.68, Student ' t '=3.1 (20) F=2.15, Fisher ' t '=1.5 (21) SNV ' z '=2.05, Fisher ' t '=2.05 (22) Fisher ' t '=5.6 (24) χ^2 =10.45 (25) χ^2 =0.28 (26) χ^2 =8.97 (27) χ^2 =0.12 (28) χ^2 =4.286 (29) χ^2 =9.48 (30) χ^2 =1.47 (31) χ^2 =39.03 (32) [a] χ^2 =0.47 (33) [a]. χ^2_T =1.241 χ^2_D =0.634, χ^2_H =0.607, χ^2 (3)=0.938, (34) χ^2 =221.52 χ^2_p =0.26, χ^2_t =0.62, χ^2_L =220.64 (35) χ^2 =4.81, χ^2_A =4.44, χ^2_B =0.018, χ^2_L =0.09 (36) χ^2 =4.5 (37) χ^2 =5.12, χ^2_A =4.90, χ^2_B =0.04, χ^2_L =0.18 (38) χ^2 =48.473, χ^2_v =0.143 χ^2_{lg} =0.073, χ^2_L =48.267 (39) χ^2 =7.084 (40) χ^2 =4.31 (41) χ^2 =0.253 (42) χ^2 =3.84 (43) χ^2 =10.989 (44) χ^2 =0.194 (45) χ^2 =4.167 (46) χ^2 =4.3.



Chapter VI

Analysis of Variance

6.1 Introduction : Generally, the planning of an experiment for some aimed at purpose is found quite troublesome, because an experiment conducted for the desired character tests some other and not only this, but sometimes it does not serve any useful purpose if not properly designed. Thus one who deals with some applied research works must be familiar with the techniques of design of experiments. One such technique of analyzing the experimental data is the technique of *analysis of variance*. It tests the homogeneity of observations with regard to the character(s) under study *i.e.* the homogeneity of several means for the aimed character(s) through the F test(s).

6.2 Meaning and definition : It is a well known fact that the variation is inherent in nature. Since we cannot find any two items exactly alike, so the variation is natural in the measurements of an experiment. The total variation present in the body of the data is generally due to several factors. These factors are of two types—

(i) Assignable factor, and (ii) Chance factor.

The assignable factor is one which is easily traceable, while the chance factor is the result of a large number of small independent causes which cannot be traced separately. This technique of analysis of variance in the first place estimates the amount of variation due to each factor separately and then compares the estimate of the assignable factor with that of the chance factor. *The estimate of the amount of variation due to the chance factor is called the experimental-error, or error.* Thus the analysis of variance can be defined as *the technique of partitioning the total variation present in the experimentally observed data into its component-variations due to the different-factors, and then comparing them.* The technique was developed by *prof. R. A. Fisher.*

6.3 Applications of analysis of variance : The analysis of variance is a powerful statistical tool* to test the homogeneity of

* It can be used as a tool to test the equality or homogeneity of several sample-means and hence may be treated as the generalization of *Fisher 't' test* used for testing the equality of two sample-means.

the observations *i.e.* to test the hyp. (H_0)—*whether the observations in the data are drawn from the same normal population*. Some other uses of this technique are made in testing the linearity of the fitted regression line and the significance of the correlation-ratio ' η '.

6.4 Assumptions for analysis of variance.

- (1) The observations are independent.
- (2) The parent-population is normal with unknown s.d. ' σ '.
- (3) The treatment and environmental-effects are additive.
- (4) The different groups are homoscedastic.

6.5 Procedure of analysis of variance : The procedure of analysis of variance with its three main sub-headings viz. *one-way classification*, *two-way classification* and *three-way classification*, is illustrated below in five main steps.

(1) Set up the hyp. (H_0)—that the data is homogeneous with regard to the factors of classification.

(2) Arrange the data in a suitable tabular-form as required according to the number of factors.

(3) Compute the total sum of squares deviated from mean and then partition it into its component-sum of squares separately.

(4) Summarize the results in the analysis of variance table called 'ANOVA'.

(5) State briefly the conclusions derived from ANOVA at some particular level of significance, say 5%.

6.5.1 One-way classification : In this type of classification, the data is classified with respect to *one factor* only. The total observations ' n ' of the data may be classified into ' k ' different groups or classes each having ' n_i ' ($i=1, 2, \dots, k$) observations with respect to some criterion of classification such that $n=\sum_i n_i$. For example, n cows of *k-different breeds* may be classified into k -classes such that n_1 of them may belong to the 1st breed, n_2 to 2nd, ..., and n_k to the k -th breed. If y_{ij} ($i=1, 2, \dots, k$ for breed no., $j=1, 2, \dots, n_i$ for cow no.) denotes the milk-yield of j th cow of i -th breed, the milk-yields of all the cows can be arranged in the following tabular-form.

Breed No.	Cow No.						Totals = T_i
	1	2	...	j	...	n_i	
1	y_{11}	y_{12}	...	y_{1j}	...	y_{1n_1}	T_1
2	y_{21}	y_{22}	...	y_{2j}	...	y_{2n_2}	T_2
...
i	y_{i1}	y_{i2}	...	y_{ij}	...	y_{in_i}	T_i
...
k	y_{k1}	y_{k2}	...	y_{kj}	...	y_{kn_k}	T_k
Total	—	—	...	—	...	—	G

In this table, T_i stands for the total milk-yield of the cows of i -th breed, and G for the grand total of the milk-yields of all the cows such that $G=\sum_i T_i=\sum_{ij} y_{ij}$.

Here we take the hyp. (H_0)-that the milk-yields are not affected by the breeds of the cows, or the breeds are not significantly different, or the data are homogeneous with regard to breeds. The factors or sources of variation are obviously—(i) breed, and (ii) error. The required sum of squares* are computed as shown below.

$$(1) \text{ Total sum of squares} = \sum_{ij} y_{ij}^2 - CF, \text{ where } CF \text{ (correction factor)} = G^2/n.$$

$$\text{i.e. TSS} = S, \text{ say.}$$

$$(2) \text{ Sum of squares due to breeds} = \sum_i T_i^2/n_i - CF$$

$$\text{i.e. SS (breeds)} = S_1, \text{ say.}$$

$$(3) \text{ Sum of squares due to error} = \text{TSS} - \text{SS (breeds)}$$

$$\text{i.e. SS (error)} = S_2, \text{ say.}$$

Now on the basis of the above mentioned hypothesis of homogeneity, the two independent estimates of the variance (σ^2) do not differ significantly. These estimates of variance are obtained by dividing the SS by their respective d.f. The significance of these estimates can be tested by using the 'F'-test. If the two estimates differ significantly at a given level, the factor of classification is said to have an influence on the variate-values. The results are summarized in the following analysis of variance table.

ANOVA

Source of variation	DF	SS	MS	F	$F_{.05}$
Breed	$k-1=v_1$	S_1	$S_1/v_1=V_1$	V_1/V_E	...
Error	$n-k=v_2$	S_2	$S_2/v_2=V_E$	—	—
Totals	$n-1$	S	—	—	—

Finally, if 'F' comes out to be significant that is, if $F \geq F_{.05}(v_1, v_2)$, we reject the hypothesis at 5% level and hence conclude that the milk-yields are affected by the breeds of the cows. But in the contrary case, we say that there is no evidence against the hypothesis at 5% level of significance.

Note : If in any case, the error MS i.e. the error variance is found to be greater than the variance of the factor, we need not to calculate the value of 'F' corresponding to the factor. Because in

* If the data are large numbers, the sum of squares may be computed from the new figures of the data obtained as residuals after deviations from some convenient assumed mean, because the SS are unaffected by the change of origin.

such a situation we simply conclude that the factor of classification is not significant since $F < 1$.

6.5.2 Two-way classification : In this type of classification, the data is classified with respect to *two factors*. The total observations 'n' of the data may be classified into 'k' different classes each having 'm' observations with respect to some criterion of classification, and into 'm' different classes each having 'k' observations with respect to some other criterion of classification such that $n = mk$. For example, n-cows of k-different breeds and m-different-lactation-periods may be classified into k-classes each having m-cows of different lactation-periods, and also into m-classes each having k-cows of different breeds. If y_{ij} ($i=1,2,\dots,k$ for breed no., $j=1,2,\dots,m$ for lactation-period no.) denotes the milk-yield of the cow of i -th breed with j -th lactation-period, then the milk-yields of all the cows can be arranged in the following tabular-form.

Breed No.	Lactation-period No.					Totals = T_i	
	1	2	...	j	... m		
1	y_{11}	y_{12}	...	y_{1j}	...	y_{1m}	T_1
2	y_{21}	y_{22}	...	y_{2j}	...	y_{2m}	T_2
...
i	y_{i1}	y_{i2}	...	y_{ij}	...	y_{im}	T_i
...
k	y_{k1}	y_{k2}	...	y_{kj}	...	y_{km}	T_k
Totals = B_j	B_1	B_2	...	B_j	...	B_m	G

In this table, T_i stands for the total milk-yield of the cows of i -th breed; B_j for the total milk-yield of the cows of j -th lactation-period; and G for the grand total of the milk-yields of all the cows such that $G = \sum_i T_i = \sum_j B_j = \sum_i \sum_j y_{ij}$.

Here we take the hyp. (H_0)—that the milk-yields are not affected, by the breeds and the lactation-periods of the cows. The sources of variation are obviously—(i) breed, (ii) lactation-period, and (iii) Error. The required sum of squares are computed as shown below—

- (1) $TSS = \sum_i y_{ij}^2 - CF = S$, say; where $CF = G^2/n$.
- (2) SS (breeds) $= \sum_i T_i^2/m - CF = S_1$, say.
- (3) SS (lact. pds.) $= \sum_j B_j^2/k - CF = S_2$, say.
- (4) And SS (error) $= TSS - SS$ (breed + lact.pd.) $= S_3$, say.

Now on the basis of the above mentioned hyp. of homogeneity, the three independent estimates of the variance (σ^2) do not differ significantly. These estimates of variance are obtained by dividing the sum of squares by their respective d.f. The significance of these estimates can be tested by using the 'F'-test. If any estimate compared with that of the error differs significantly at a give level, the corresponding factor of classification is said to have an

influence on the variate-values. The results are summarized in the following analysis of variance table.

ANOVA

Source of variation	DF	SS	MS	F	F _{0.05}
Breed	$k-1=v_1$	S_1	$S_1/v_1=V_1$	V_1/V_E	...
Lact. pd.	$m-1=v_2$	S_2	$S_2/v_2=V_2$	V_2/V_E	...
Error	$(k-1)(m-1)=v_3$	S_3	$S_3/v_3=V_E$	---	---
Totals	$n-1$	S	---	---	---

Finally, if both or any one 'F' comes out to be significant, we reject the hyp. at 5% level corresponding to both or any one factor; otherwise we conclude that there is no evidence against the hyp. at 5% level of significance.

6.5.3 Three-way classification : In this type of classification, the data is classified with respect to *three factors*. The total observations 'n' of the data may be classified into 'k' different classes with respect to some one criterion of classification, into 'm' different classess with respect to some other criterion, and into 'p' different classes with respect to the remaining third criterion of classification. For example, n-cows of k-different breeds, m-different lactation-periods and p-different age-groups may be classified into k, m, p classes respectively according to the said factors of classification. But here we shall consider the particular case where the no. of classes with respect to each of the three factors is the same i.e. $k=m=p$. This type of arrangement is done in Latin Squares, where the rows denote the classes with respect to first factor, columns with respect to second factor and the latin-letters denote the classes with respect to the third factor of classification. In a latin square, the observations are so arranged that each letter occurs only once in each row and each column. Also, a latin square with k-rows and k-columns is called a ' $k \times k$ ' Latin Square, or a Latin-Square of order 'k'. If for the above said example, we take for convenience $k=3$, then the milk-yields denoted by y_{ij} ($i=1, 2 \dots k$ for breed no. $j=1, 2, \dots k$, for lactation-period no.), along with a latin letter can be arranged in the following tabular-form.

Breed No.	Lact. pd. no. and age grp. No.			Totals = R_i
	1	2	3	
1	A	B	C	R_1
2	y_{11}	y_{12}	y_{13}	R_2
3	y_{21}	y_{22}	y_{23}	R_3
	C	A	B	
	y_{31}	y_{32}	y_{33}	
Totals = C_j	C_1	C_2	C_3	G
Age grp. Totals = T_a	T_a	T_b	T_c	
	

In this table, R_i stands for the total milk-yield of i -th row corresponding to i -th breed; C_j for the total milk-yield of j -th column corresponding to j -th lactation period; T_a for the total milk-yield of the cows of 'A'-age-group; and G for the grand total of the milk-yields of all the cows such that $G = \sum_i R_i = \sum_j C_j = \sum_a T_a = \sum_{ij} y_{ij}$, $n = k^2$.

It is quite clear that the symbol $\left(\begin{smallmatrix} A \\ y_{11} \end{smallmatrix} \right)$ represents the milk-yield

of the cow which is of breed no.1, lactation-period no.1 and age-group 'A'. The similar meanings are attached with the remaining symbols used in the table.

Here we take the hyp. (H_0)—that the milk-yields are not affected by the breeds, lactation-periods and the age-groups of the cows. The sources of variation are obviously—(i) breed, (ii) lactation period, (iii) age-group, and (iv) error. The required sum of squares are computed as shown below—

(1) $TSS = \sum_{ij} y_{ij}^2$, $CF = S$, say; where $CF = G^2/n$.

(2) SS (breeds, or rows) $= \sum_i R_i^2/k - CF = S_1$, say.

(3) SS (lact. pds., or cols.) $= \sum_j C_j^2/k - CF = S_2$, say.

(4) SS (age grps., or letters) $= \sum_a T_a^2/k - CF = S_3$, say.

(5) SS (error) $= TSS - SS$ (rows + cols. + letters) $= S_4$, say.

Now on the basis of the above mentioned hyp. of homogeneity, the four independent estimates of the variance (σ^2) do not differ significantly. These estimates of variance are obtained by dividing the sum of squares by their respective degrees of freedom. The significance of these estimates can be tested by using the 'F'-test. If any estimate compared with that of the error differs significantly at a given level, the corresponding factor of classification is said to have an influence on the variate-values. The results are summarized in the following analysis of variance table.

ANOVA

Source of variation	DF	SS	MS	F	$F_{.05}$
Breed	$k-1 = v_1$	S_1	$S_1/v_1 = V_1$	V_1/V_E	...
Lact. pd.	$k-1 = v_1$	S_2	$S_2/v_1 = V_2$	V_2/V_E	...
Age. grp.	$k-1 = v_1$	S_3	$S_3/v_1 = V_3$	V_3/V_E	...
Error	$(k-1)(k-2) = v_2$	S_4	$S_4/v_2 = V_E$	—	—
Totals	$n-1$	S	—	—	—

Finally, if one or more values of 'F' come out to be significant, we reject the hyp. at 5% level; otherwise we conclude that there is no evidence against the hyp. at 5% level of significance.

Exp. (1) For the following data, prepare the analysis of variance table and test the significance of the difference between the yields of the three varieties.

variety	yields in lbs.				
A :	10	12	8	10	
B :	13	10	11		
C :	9	12	13	14	10.

Sol. (H_0) : The three varieties A, B and C are not significantly different as regards their yielding capacities.

If y_{ij} ($i=1, 2, 3$ for variety no., $j=1, 2, \dots, n_i$ for plot no.) denotes the yield (in lbs.) of j -th plot for i -th variety, the yields of all the plots can be arranged in the following (one-way classification) tabular form.

Variety No.	PLOT No.					Totals $=T_i$	No we compute— $CF = G^2/n$ $= (132)^2/12$ $= 1452.$
	1	2	3	4	5		
1. A	10	12	8	10	—	40	
2. B	13	10	11	—	—	34	
3. C	9	12	13	14	10	58	

$$TSS = \sum y_{ij}^2 - CF = 1488 - 1452 = 36,$$

$$SS \text{ (varieties)} = \sum T_i^2/n_i - CF$$

$$= (40)^2/4 + (34)^2/3 + (58)^2/5 - 1452 = 1458.13 - 1452 = 6.13,$$

$$SS \text{ (error)} = TSS - SS \text{ (varieties)} = 36 - 6.13 = 29.87.$$

ANOVA

Source of variation	DF	SS	MS	F	$F_{.05}$
Variety	2	6.13	3.065	< 1	4.26
Error	9	29.87	3.32	—	—
Totals	11	36.00	—	—	—

Conclusion : Since ' F ' < $F_{.05}$ (2,9) showing its insignificance, so there is no evidence against the hyp. and we conclude that the varieties do not differ significantly as regards their yielding capacities.

Exp. (2) A certain company had four salesmen, A, B, C and D each of whom was sent for a week into three types of areas—country area 'K', outskirts of city 'O' and shopping centre of a city 'S'. The sales in pounds per week are shown below.

Area **Sales in pounds/week**

	A	B	C	D
K :	30	70	30	30
O :	80	50	40	70
S :	100	60	80	80.

Garry out an analysis of

variance and interpret the result stating the limitations under which your conclusions are valid ? (ICAR, Delhi 1956)

Sol. (H_0): *There is no significant difference between the sales of the four salesmen and the three types of areas.*

Since the figures of the data are large numbers, so we can use the deviations from some convenient assumed mean, say $y=60$. If y_{ij} ($i=1, 2, 3$ for area no., $j=1, 2, 3, 4$ for salesman no.) denotes the sale (in pounds/week) of j th salesman for i th area, then the sales of all the places can be arranged in the following (two-way classification) tabular form.

Area No.	SALESMAN No.				Totals $=T_i$
	1.A	2.B	3.C	4.D	
1.K	-30	10	-30	-30	-80
2.O	20	-10	-20	10	0
3.S	40	0	20	20	80
Total $=B_j$	30	0	-30	0	0=G

Now we compute—

$$CF = G^2/n \\ = (0)^2/12 \\ = 0.$$

$$TSS = \sum_i \sum_j y_{ij}^2 - CF = 6200 - 0 = 6200,$$

$$SS \text{ (areas)} = \sum_i T_i^2/n - CF = \frac{(-80)^2 + (0)^2 + (80)^2}{4} - 0 = 3200$$

$$SS \text{ (salesmen)} = \sum_j B_j^2/k - CF = \frac{(30)^2 + (0)^2 + (-30)^2 + (0)^2}{3} - 0 = 600,$$

$$SS \text{ (error)} = TSS - SS \text{ (areas + salesmen)} = 2400.$$

ANOVA

Source of variation	DF	SS	MS	F	$F_{.05}$
Area	2	3200	1600	$1600/400=4$	5.14
Salesman	3	600	200	$200/400 < 1$	8.94
Error	6	2400	400	—	—
Totals	11	6200	—	—	—

Conclusion : The insignificant values of F , show that there is no evidence against the hyp. at 5% level, and thus we conclude that there is no significant difference between the sales of the four salesmen and three types of areas.

Exp. (3) In an experiment on the spacing of millet, four spacings were used— $A=2''$, $B=4''$, $C=6''$ and $D=8''$, and yields were arranged in a latin square. The experimental arrangement with yields in grams/plot is shown below—

B	D	A	C
249	245	249	244
A	C	D	B
254	249	240	252
D	B	C	A
245	254	250	257
C	A	B	D
251	261	254	246.

Construct an analysis of variance and test for the variation between the spacings?

Sol. (H_0) : *There is no significant difference between the yields of the plots due to four rows, columns and spacings.*

Since the figures of the data are large numbers, so we can use the deviations from some convenient assumed mean, say $y=250$. If y_{ij} ($i=1, 2, 3, 4$ for row no., $j=1, 2, 3, 4$ for col. no.) denotes the yield (in gms/plot) of j -th column for i -th row, the letters A, B, C and D stand for spacings, then the yields of all the plots can be arranged in the following (*three-way classification*) tabular form.

Row No.	Col. no. and spacing No.				Totals $=R_i$
	1	2	3	4	
1	B	D	A	C	
	-1	-5	-1	-6	-13
2	A	C	D	B	
	4	-1	-10	2	-5
3	D	B	C	A	
	-5	4	0	7	6
4	C	A	B	D	
	1	11	4	-4	12
Totals $=C_j$	-1	9	-7	-1	0=G
Spacing Totals $=T_a$	T_a	T_b	T_c	T_d	
	21	9	-6	-24	

Now we compute—

$$CF = G^2/n = (0)^2/16 = 0$$

$$TSS = \sum_i \sum_j y_{ij}^2 - CF = 428$$

$$SS(\text{rows}) = \sum_i R_i^2/k - CF = 93.5$$

$$SS(\text{cols.}) = \sum_j C_j^2/k - CF = 33.0$$

$$SS(\text{spacings}) = \sum_a T_a^2/k - CF = 283.5,$$

$$SS(\text{error}) = TSS - SS(\text{rows} + \text{cols.} + \text{spacings}) = 428 - (93.5 + 33.0 + 283.5) = 18.$$

ANOVA

Source of variation	DF	SS	MS	F	F _{.05}
Row	3	93.5	31.17	10.39	4.76
Column	3	33.0	11.04	3.67	„
Spacing	3	283.5	94.50	31.50	„
Error	6	18.0	3.00	—	—
Totals	15	428.0	—	—	—

Conclusion : The significant values of F , corresponding to rows and spacings indicate that there are significant differences

between the yields of the plots due to the row-means and the spacing-means.

Exercise VI

1. To test the significance of variation of the retail prices of a certain commodity in the four principal cities—Bombay, Calcutta, Madras and Delhi, seven shops were chosen at random in each city and the prices observed were as follows—

City *Prices in np. at the shops*

Bombay : 82 79 73 69 69 63 61

Calcutta : 84 82 80 79 76 68 62

Madras : 79 77 76 74 72 68 64

Delhi : 88 84 80 68 68 66 66. Do the data indicate that the prices in the four cities are significantly different? Tabulate the results properly for the study of variation both between cities and within cities?

2. The plants of wheat of four varieties were selected at random and the heights of their shoots were measured in cms.—

Variety *Heights in cms.*

1 : 40 42 41 43 45

2 : 42 41 44 45 43

3 : 39 42 43 40 44

4 : 37 34 38 35 37. Do the data indicate that there is no significant difference between the mean heights of the plants of the four varieties?

3. Four varieties of potato are planted, each on five plots of ground of the same size and type; and each variety is treated with five different fertilizers. The yields in tons/plot are as follows—

Fertilizers

variety 1 2 3 4 5

1 : 1.9 2.2 2.6 1.8 2.1

2 : 2.5 1.9 2.3 2.6 2.2

3 : 1.7 1.9 2.2 2.0 2.1

4 : 2.1 1.8 2.5 2.3 2.4. Perform an analysis of variance and show whether there is any significant difference between the yields of four varieties or due to five fertilizers.

4. [a] Define “analysis of variance”. Give its assumptions and also the applications?

[b] Six varieties of wheat were tested in the four blocks of a field. The yields in kgms/plot and the layout are given below.

Block	Layout and yields/plot					
I :	V ₁	V ₃	V ₂	V ₄	V ₅	V ₆
	17.8	17.7	20.6	6.2	6.2	14.9
II :	V ₃	V ₂	V ₁	V ₄	V ₆	V ₅
	12.7	18.8	17.3	5.0	12.5	7.0
III :	V ₆	V ₄	V ₁	V ₃	V ₂	V ₅
	16.3	9.6	28.5	26.8	29.5	5.4
IV :	V ₅	V ₂	V ₁	V ₄	V ₃	V ₆
	7.7	21.0	18.5	4.1	24.9	12.6.

Prepare the analysis of variance table and test whether the varieties and the blocks differ significantly in regard to their average-yields ?

5. You are given the results of the following cacao Manurial experiment conducted in a 3×3 latin square. The fertilizers used were—O→no manure (control), A→one lb. of super phosphate/plant and B →two lbs. of super phosphate/plant. The layout and yields in lbs./plant from $\frac{1}{10}$ acre plots are given below—

O	A	B
14	40	24
B	O	A
23	19	31
A	B	O
20	21	

11. Carry out the analysis of variance and test the variation between the fertilizers.

6. Below are given the yields in seers/plot, and the plan of four varieties of gram tested in a 4×4 latin square.

A	B	C	D
105	112	113.1	108
B	D	A	C
113.5	110.5	114	112
C	A	D	B
114	108.5	110	113
D	C	B	A
107.5	115.5	111.1	111.5.

Carry out the analysis of variance and test the homogeneity of the given data ?

ANSWERS

(1) $V_1 = 31.67$, $V_E = 60.25$ (2) $F = 13.96$ (3) $V_1 = 0.0953$, $V_2 = 0.1156$, $V_E = 0.0568$ (4) [b]. $F_1 = 21.98$, $F_2 = 5.77$ (5) $F_1 = 10.7$, $F_2 = 5.9$, $F_3 = 4.15$ (6) $V_R = 6.09$, $V_C = 3.13$, $V_T = 19.18$, $V_E = 5.68$.



Chapter VI1

Correlation And Regression

7.1 Meanings and definition of correlation : A number of statistical problems arise in exact and social sciences in which the sample drawn from a bivariate normal population consists of pairs of measurements (x, y) . In all such problems, these two variables (x, y) are found to behave in such a manner that the change in one brings the change in the other. This type of phenomenon or relationship between any two variables is called *correlation* and such variables are said to be *correlated*. Thus the correlation may be defined as *the relationship between the two variables when the change in one variable is on an average accompanied by the change in the other variable in the same or opposite directions*. For example, the two variables-volume and pressure of a perfect gas at some constant temperature, according to Boyles' Law, are so related that the volume increases with the decrease in pressure or vice versa, and hence these two variables are said to have the correlation.

7.1.1 Direction of correlation : By direction of correlation we mean the *sign* of correlation. On the basis of the direction, the correlation is of the following two types.

(i) Positive correlation :

The correlation between the two variables is said to be (+)ve if the changes in both the variables

Direction of correlation

- (i) Positive correlation.
- (ii) Negative correlation.

are observed in the same direction. It clearly means that either both the variables increase or decrease simultaneously. This type of correlation is found between—the total cultivable area and the area under wheat; the amount of production of a crop and the amount of fertilizer applied to it; the demand and the price of a certain commodity; the age and the height of a child; the radius and the circumference or area of a circle; the volume and the temperature of a perfect gas at some constant pressure according to Charle's law etc.

(ii) Negative correlation : The correlation between the two variables is said to be (—)ve if the changes in the two variables

are observed in opposite directions. It clearly means that if one variable is increasing, the other is decreasing, and vice versa. This type of correlation is found between—the areas under fodder and grain-crops; the supply and the price of a certain commodity, the total income and the proportion of it spent on food; the volume and the pressure of a perfect gas at some constant temperature according to Boyle's Law etc.

7.1.2 Degree of correlation : By degree of correlation we mean the magnitude or extent of correlation. On the basis of the degree *i.e.* the ratio of changes in the two variables; the correlation is of the following two types.

Degree of correlation

- (i) Perfect correlation.
- (ii) Limited correlation.

(i) Perfect correlation : The correlation between the two variables is said to be perfect if the ratio of changes in both the

variables remains the same throughout. It clearly means that the percentage-change in one variable is accompanied by the same percentage-change in the other variable in the same or opposite directions. This type of correlation is found between—the radius and the circumference of a circle; the volume and the pressure of a perfect gas at some constant temperature etc.

(ii) Limited correlation : The correlation between the two variables is said to be limited if the ratio of changes in the two variables does not remain the same throughout. It clearly means that the percentage-change in one variable is not accompanied by the same percentage-change in the other variable in the same or opposite directions. This type of correlation is found between—the demand and the price of a certain commodity; the areas under fodder and grain crops etc.

7.1.3 Degree and direction of correlation : By degree and direction of correlation we mean both—the *magnitude* (or extent) and the *sign* of correlation. On the basis of the degree as well as the direction, the correlation is of the following four types.

Degree and direction of correlation

- (i) Perfect positive correlation
- (ii) Limited positive correlation
- (iii) Perfect negative correlation
- (iv) Limited negative correlation

(i) Perfect positive correlation: The correlation between the two variables is said to be perfect positive if the direction and the ratio of changes in both the variables remain the same throughout. This type of

correlation is found between—the volume and the temperature of a perfect gas at some constant pressure.

(ii) **Limited positive correlation** : The correlation between the two variables is said to be limited positive if the direction of changes in the two variables remains the same throughout but the ratio of changes differs. This type of correlation is found between the total cultivable area and the area under wheat.

(iii) **Perfect negative correlation** : The correlation between the two variables is said to be perfect negative if the direction of changes in the two variables differs but the ratio of changes remains the same throughout. This type of correlation is found between—the volume and the pressure of a perfect gas at some constant temperature.

(iv) **Limited negative correlation** : The correlation between the two variables is said to be limited negative if both—the direction and the ratio of changes in the two variables differ throughout. This type of correlation is found between—the areas under fodder and grain-crops.

7.1.4 No correlation : The two variables are said to have no correlation (or uncorrelated) if the change in one variable does not affect the other variable. For example, the *no. of radio sets produced* and the *total no. of births recorded* during a certain period have no relationship with each other, and hence the two have no correlation.

7.2 Measures of correlation : The direction and the magnitude of correlation between the two variables can be found by using either one of the *mathematical methods* or a *diagrammatic one*.

Mathematical methods

7.2.1 Mathematical methods.

7.2.1 (a) Pearson's coefficient of correlation : This measure of correlation obtained by *prof. Karl Pearson* is based on the

- (a) Pearson's coefficient of correlation.
- (b) Spearman's „ „ „
- (c) Coefficient of concurrent deviations.
- (d) Coefficient of correlation by least-squares.

arithmetical descriptions. It is usually denoted by the symbol ' r '. This measure is quite capable of expressing the direction and the exact amount of casual relationship between the two variables under study. It has been seen that the values of ' r ' always lie between -1 and $+1$. Further we note that for a value $r = +1$, there is perfect (+) ve correlation; for $r = -1$, there is perfect (-) ve

correlation; and for $r = 0$, there is no correlation between the two variables. Also for a value of r lying between 0 and 1, there is limited (+) ve correlation while for r lying between 0 and -1, there is limited (-)ve correlation between the two variables. If we consider a random sample of n pairs in (x, y) drawn from a bivariate normal population, then Pearson's coefficient of correlation or the product moment coefficient of correlation is given by the formula—

$$r = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y}, \text{ where } \text{cov}(x, y) = \Sigma (x - \bar{x})(y - \bar{y})/n \text{ stands}$$

for the sample-covariance of x and y , $\sigma_x = \sqrt{[\Sigma (x - \bar{x})^2/n]}$ and $\sigma_y = \sqrt{[\Sigma (y - \bar{y})^2/n]}$ stand for sample-s.d.s. of x and y respectively. Thus by using the *direct method*, we have

$$r = \frac{\Sigma (x - \bar{x})(y - \bar{y})}{\sqrt{[\Sigma (x - \bar{x})^2 \Sigma (y - \bar{y})^2]}}$$

For the sake of convenience in computations especially when the actual means \bar{x} , \bar{y} come out to be in decimals, we can also use a *short cut method* to compute ' r ' as

$$r = \frac{\Sigma \xi \eta - \Sigma \xi \Sigma \eta / n}{\sqrt{[\Sigma \xi^2 - (\Sigma \xi)^2 / n][\Sigma \eta^2 - (\Sigma \eta)^2 / n]}} \text{, where } \xi, \eta \text{ stand for deviations}$$

or step-deviations of the variates x and y from their respective assumed means A_x and A_y . This method is very common in practice and it consists of the following main steps.

- (i) Prepare a blank table with eight columns in it.
- (ii) Use I col. for *pair no.*; II—for x ; III—for the deviation $\xi = (x - A_x)$ or the step-deviation $\xi = (x - A_x)/i$ if i be the common factor; IV—for ξ^2 ; V—for y ; VI—for the deviation $\eta = (y - A_y)$ or the step-deviation $\eta = (y - A_y)/i$ if i be the common factor; VII—for η^2 and VIII—for the product $\xi \eta$.

(iii) Complete the entries of all the eight columns against each pair number.

(iv) In the bottom of the table, get the totals of all the cols. except the col. no. I, II, V, and substitute them in the formula to compute the value of ' r '.

7.2.1 (a-1) Assumptions for Pearson's coefficient of correlation :

- (i) The two variables are linearly related.
- (ii) The variables are affected by a no. of independent causes.
- (iii) The independent causes, affecting the variables, have some inter-relationship between the mutual causes and effects.
- (iv) The variables are random.

(a-2) Properties of Pearson's coefficient of correlation :

- (i) It gives us the direction as well as the magnitude of correlation between the two variables.
- (ii) Its value always lies between -1 and $+1$.
- (iii) It is independent of the change of origin and scale.
- (iv) It is symmetrical in the two variables, i.e. $r_{xy} = r_{yx}$.
- (v) It is rigidly defined and hence free from human bias.
- (vi) It depends upon all the observations of the data.
- (vii) It is a pure number and hence free from any of the units.
- (viii) It bears the same sign as that of $\text{cov}(x, y)$.

(a-3) Limitations of Pearson's coefficient of correlation :

- (i) It is difficult to compute.
- (ii) Its meanings are not easily understood.
- (iii) It does not show whether there is any relation between the causes and effects producing the correlation.

(a-4) Applications of Pearson's coefficient of correlation :

- (i) It can be used to find the relation between the two variables.
- (ii) It can be used to determine the two regression coefficients, the angle between the two regression lines, the standard errors of estimates, and the covariance between the two variables provided the s.d.s. of the variables are known.
- (iii) It can also be used as a measure of linearity between the two variables.

Exp.(1) Calculate the coefficient of correlation for the following ages of husband and wife—

Husband's age (x yrs.)...	23	27	28	29	30	31	33	35	36	39
Wife's age (y yrs.)...	18	22	23	24	25	26	28	29	30	32

Sol. The following table shows the calculation of r .

Pair No.	x	$\xi = x - A_x$ $A_x = 31$	ξ^2	y	$\eta = y - A_y$ $A_y = 25$	η^2	$\xi\eta$
1	23	-8	64	18	-7	49	56
2	27	-4	16	22	-3	9	12
3	28	-3	9	23	-2	4	6
4	29	-2	4	24	-1	1	2
5	30	-1	1	25	0	0	0
6	31	0	0	26	1	1	0
7	33	2	4	28	3	9	6
8	35	4	16	29	4	16	16
9	36	5	25	30	5	25	25
10	39	8	64	32	7	49	56
Totals	—	$1 = \sum \xi$	$203 = \sum \xi^2$	—	$1 = \sum \eta$	$163 = \sum \eta^2$	$179 = \sum \xi\eta$

Now we have—

$$\begin{aligned}
 r &= \frac{\Sigma \xi \eta - \Sigma \xi \Sigma \eta / n}{\sqrt{[\{\Sigma \xi^2 - (\Sigma \xi)^2 / n\} \{\Sigma \eta^2 - (\Sigma \eta)^2 / n\}]}} \\
 &= \frac{179 - 1 \times 7 / 10}{\sqrt{[\{203 - (1)^2 / 10\} \{163 - (7)^2 / 10\}]} } = \frac{179 - 7}{\sqrt{[\{203 - 1\} \{163 - 4.9\}]} } \\
 &= \frac{178.3}{178.6} = +.90.
 \end{aligned}$$

Thus there is (+) ve correlation of high degree between the ages of husband and wife. **Ans.**

(b) Spearman's coefficient of correlation : This measure of correlation, obtained by *Spearman Brown*, is also based on the arithmetical descriptions. *It gives the correlation between the ranks (or grades) assigned to the two characters under study and hence may also be called as the rank correlation coefficient.* It is computed in the situations when the numerical measurements on the characters are difficult but their grading is easy with regard to some criterion; or when we want to know the relationship between the proficiencies of a group of candidates in the two different subjects, or when we want to investigate the degree of agreement between the two judges who have graded the same individuals regarding the same characteristic.

It is also denoted by the symbol r . This measure is also capable of expressing the direction and the exact amount of casual relationship between the two characters under study. Its value also lies between -1 and $+1$. *If we consider a random sample of n pairs in (x, y) drawn from a bivariate normal population, then Spearman's coefficient of correlation is given by the formula—*

$$r = 1 - \frac{6 \Sigma d^2}{n(n^2 - 1)}, \text{ where } d \text{ stands for the difference between the}$$

two ranks of the same individual. Here it is assumed that no two or more individuals receive the same rank for a character. If two or more individuals possess the same magnitude for a character, they receive the average rank determined on the basis as if their magnitudes are slightly different from each other. An individual possessing the highest magnitude of a character is usually assigned rank 1; the next lower rank 2; and so on,, the individual with lowest magnitude is assigned the highest rank for the character which is the same as n , the no. of pairs in the sample. This method of finding the rank correlation coefficient between the two characters consists of the following main steps—

(i) Prepare a blank table with seven columns in it.

(ii) Use I-col. for pair no., II-for x , III-for rank of x ; IV-for y ; V-for rank of y ; VI-for rank difference 'd'; and VII-for d^2 .

(iii) Complete the entries of all the seven columns against each pair number.

(iv) In the bottom of the table, get the total of the last col. only and substitute it in the formula to compute the value of r .

Note : If in a problem, the direct ranks are given for the individuals of the sample, then we need not to include the cols. II, IV in the above table. The assumptions, properties, limitations and the applications of the rank correlation coefficient are the same as those stated for Pearson's coefficient of correlation.

Exp (2) Ten students got the following percentage of marks in Economics and Statistics —

Student ... 1 2 3 4 5 6 7 8 9 10

Marks in Eco. ... 78 36 98 25 75 82 90 62 65 39

Marks in Stat ... 84 51 82 45 82 62 82 58 53 47. Calculate the coefficient of rank correlation.

Sol. The computation of rank correlation coefficient is shown in the following table:

Student No.	Marks in Eco. x	Rank of x	Marks in Stat. y	Rank of y	Rank diff. d	d^2
1	78	4	84	1	3	9
2	36	9	51	8	1	1
3	98	1	82	3	-2	4
4	25	10	45	10	0	0
5	75	5	82	3	2	4
6	82	3	62	5	-2	4
7	90	2	82	3	-1	1
8	62	7	58	6	1	1
9	65	6	53	7	-1	1
10	39	8	47	9	-1	1

Now we have

$$r = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

$$= 1 - \frac{6 \times 26}{10(100-1)} = 1 - \frac{156}{990} = 1 - .16 = +.84.$$

Thus there is high (+)ve correlation between the marks obtained in Economics and Statistics. Here we may also conclude that the students who are good in Eco. are also good in Statistics.

... Ans.

7.2.2 Diagrammatic methods :

(a) **Scatter diagram** : This measure of correlation does not require any arithmetical descriptions but is simply based on the nature of the diagram obtained on a graph paper. If we consider a random sample of n pairs in (x, y) drawn from a bivariate normal population and plot these points on a graph paper, then the diagram of dots (or points) thus obtained is called scatter diagram.

Diagrammatic methods

- (a) Scatter diagram.
- (b) Regression lines.
- (c) Simple graphs.

Merely an eye-inspection of this diagram is sufficient to decide the degree and direction of correlation between the two variables.

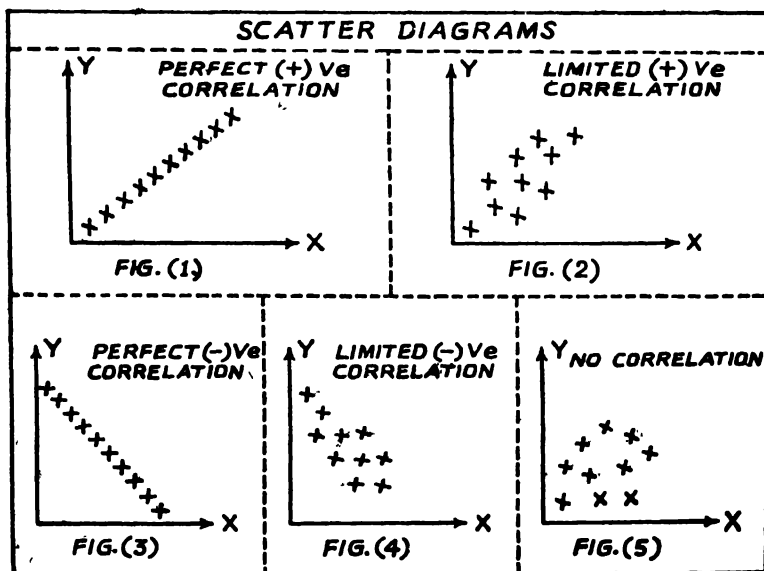
(1) If all the points on a scatter diagram lie on a straight line with (+) ve slope, the correlation will be *perfect (+) ve* [see fig. (1)].

(2) If the points on a scatter diagram seem to cluster about the diagonal with (+)ve slope, the correlation will be *limited (+) ve* [see fig. (2)].

(3) If all the points on a scatter diagram lie on a straight line with (—) ve slope, the correlation will be *perfect (—)ve* [see fig. (3)].

(4) If the points on a scatter diagram seem to cluster about the diagonal with (—)ve slope, the correlation will be *limited (—)ve* [see fig. (4)].

(5) If the points on a scatter diagram take roughly the elliptical or circular shape, then the two variables are said to be *uncorrelated* [see fig. (5)].



(a-1) Properties of scatter diagram :

- (i) It gives us the direction as well as the degree of correlation between the two variables.
- (ii) It depends upon all the observations of the data.
- (iii) It is simply a graph and hence free from any arithmetical descriptions.
- (iv) It is a very quick measure of correlation.
- (v) It is easy to understand.

(a-2) Limitations of scatter diagram :

- (i) It does not give us the exact amount of correlation but merely the degree of correlation *i.e.* perfect or limited.
- (ii) It is not rigidly defined but depends upon the accuracy and skill of the observer and hence not free from human bias.
- (iii) It can be used only for elementary purposes where we need simply a rough idea of correlation between the two variables.
- (iv) It does not show whether there is any relation between the causes and effects producing the correlation.
- (v) It is unable to give us any idea of correlation for a small sample because no specific diagram can be obtained from these few points.

(a-3) Applications of scatter diagram :

- (i) It can be used to find the relation between the two variables.
- (ii) It can be used as a pictogram for advertisement purposes to illustrate the relationship between the two characters.
- (iii) It can be used as a quick measure of correlation in all those situations wherein we require simply the approximate idea of correlation.

Note : *The assumptions for a scatter diagram are the same as described for Pearson's coefficient of correlation.*

7.2.2(b) Regression lines : The coefficient of correlation between the two variables can also be found with the help of the two regression lines. Let us consider a random sample of n pairs in (x, y) drawn from a bivariate normal population and plot these points on a graph paper to have the scatter diagram. *If there exists association or relationship between the two variables x and y , the dots of the scatter diagram may be more or less concentrated around a line called the line of regression and the relationship thus exhibited is called the linear regression.* Thus a regression line may also be termed as a regression function of order one. More precisely, *a line of regression is the straight line which gives the*

best fit in the least square sense, to the given frequency distribution.

If a straight line is so chosen that the sum of squares of deviations parallel to the axis of x is minimum, it is called *the line of regression of x on y* and it gives the best estimate of x for a given value of y . Similarly, if the sum of squares of deviations parallel to the axis of y is minimized, the resulting straight line is called *the line of regression of y on x* and it gives the best estimate of y for a given value of x . Usually, these two regression lines are different.

Thus the regression concepts concerned with the way in which the changes in one variable depend upon the simultaneous changes in the other variable. Out of the two random variables used for correlation theory, here we choose one as the independent variable while the other as dependent. The variable whose effect produces the correlation is treated as independent and the other as dependent. For example, the correlation between *the amount of rainfall* and *yield of rice* is produced due to the influence of rainfall on the yield of rice and hence the former is treated as the independent variable, the latter as dependent. Sometimes both the variables may take both of the roles. It happens when the correlation between them is the result of the influence of a third factor. For example, the (+)ve correlation between *the yields of rice and jute* is due to the fact that the two are related to the amount of rainfall. Hence we see that in a linear regression there may be at the most two regression lines as

$$(i) y - \bar{y} = b_{yx}(x - \bar{x}), \text{ and } (ii) x - \bar{x} = b_{xy}(y - \bar{y}).$$

In the above relations, (i) represents the equation of regression line of y on x which can give the estimated (or average, or expected or most likely) value of the dependent variable y for a given value of the independent variable x . Similarly, (ii) represents the equation of regression line of x on y which can give the best estimated value of x for a given value of y . Hence \bar{x}, \bar{y} denote the sample means for the variables x, y respectively and b_{yx}, b_{xy} stand for the respective regression coefficients of y on x and of x on y . *The regression coefficient b_{yx} gives the average change in y for a unit change in x .* In the same way, the regression coefficient b_{xy} can be defined as the average change in x for a unit change in y . These coefficients of regression can be determined by any of the following formulae—

$$(i) b_{yx} = \frac{\text{cov}(x, y)}{\sigma_x^2},$$

$$(ii) b_{yx} = \frac{r\sigma_y}{\sigma_x},$$

$$(iii) b_{yx} = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2},$$

$$\text{and } (iv) b_{yx} = \frac{\sum \xi \eta - \sum \xi \sum \eta / n}{\sum \xi^2 - (\sum \xi)^2 / n},$$

where \bar{x}, \bar{y} stand for the sample means; σ_x, σ_y for the sample s.ds., and ξ, η for the deviations of the variables x and y from their respective assumed means A_x and A_y . similarly, we may have

$$\begin{aligned} \text{(i) } b_{xy} &= \frac{\text{cov}(x, y)}{\sigma_y^2}, & \text{(ii) } b_{yx} &= \frac{r\sigma_x}{\sigma_y}, \\ \text{(iii) } b_{xy} &= \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(y - \bar{y})^2}, & \text{and (iv) } b_{yx} &= \frac{\Sigma\xi\eta - \Sigma\xi\Sigma\eta/n}{\Sigma\eta^2 - (\Sigma\eta)^2/n}. \end{aligned}$$

For computing a regression coefficient by the last formula, the procedure consists of the same steps as stated for Pearson's coefficient of correlation.

The direction and the magnitude of correlation between the two variables can be determined with the help of the two regression lines by using either (a) *the two regression coefficients*, or (b) *the angle between the two regression lines*.

7.2.2 (b-a) Determination of r from two regression coefficients : Given the two regression coefficient, the correlation coefficient can be determined as the geometric mean of the regression coefficients. The sign of r is the same as that of either of the two regression coefficients because b_{yx}, b_{xy}, r and $\text{cov}(x, y)$ bear the same sign since σ_x, σ_y are always (+) ve.

$$\text{Thus we have} \quad b_{yx} \times b_{xy} = \frac{r\sigma_y}{\sigma_x} \times \frac{r\sigma_x}{\sigma_y} = r^2 (\leq 1)$$

$$\text{or} \quad \pm(b_{yx} \cdot b_{xy})^{1/2} = r.$$

(b.a-1) Properties of regression coefficients :

(i) They give us the estimates of changes of the dependent variables corresponding to unit changes in the independent variables.

(ii) One of the two regression coefficients may be less than unity and the other greater than unity, but their product must never exceed unity.

(iii) They are independent of the change of origin, but not of scale.

(iv) They are not symmetrical in the two variables, i.e. $b_{yx} \neq b_{xy}$ in general.

(v) They are rigidly defined and hence free from human bias.

(vi) They depend upon all the observations of the data.

(vii) They are the pure numbers and hence free from any of the units.

(viii) They bear the same sign as that of the correlation, or the covariance term.

(b.a-2) Applications of regression coefficients :

(i) A regression coefficient can be used to find the amount of

correlation between the two variables x and y provided the sample s.d.s σ_x , σ_y are known.

(ii) A regression coefficient $b_{y/x}$ (or b_{xy}) can also be used to find the covariance between the two variables x and y provided the sample s.d. σ_x (or σ_y) is known.

(iii) The two regression coefficients together can be used to determine the coefficient of correlation, the slopes of the regression lines with the coordinate axes, the angle between the two regression lines, the standard errors of estimates and the covariance between the two variables provided the sample s.d.s, σ_x , σ_y are known.

(iv) Their product ($=r^2$) can be used as a measure of linearity between the two variables.

(v) They are also used to determine the two regression equations (or functions) provided the sample means \bar{x} , \bar{y} are known.

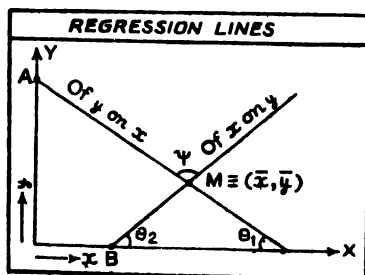
Note : *The assumptions and the limitations of the regression coefficients are the same as stated for Pearson's coefficient of correlation.*

7.2.2 (b-b) Determination of r from the angle between the two regression lines : If the angle between the two regression lines plotted on a graph is ψ , its magnitude can be used to determine the amount of correlation between the two variables under study.

If $\psi=0$, the correlation is perfect (+)ve; if $0<\psi<90^\circ$, the correlation is limited (+)ve; if $\psi=90^\circ$, there is no correlation at all, if $90^\circ<\psi<180^\circ$, the correlation is limited (-)ve; and if $\psi=180^\circ$, the correlation is perfect (-)ve.

(b.b-1) Plotting the two regression lines on a graph :

Although for plotting the two lines on a graph we need at least four points, but here only three points are sufficient since a point $M \equiv (\bar{x}, \bar{y})$ is common to both the regression lines being their point of intersection. Thus we need to plot the other two points, say A and B. If we put $x=0$ in the equation of y on x and find the corresponding value of y , say y_0 , then we obtain the point $A \equiv (0, y_0)$. If we join M and A, we get the line of regression of y on x . Similarly,



by putting $y=0$ in the equation of x on y and finding the corresponding value of x , say x_0 , obtain the point $B \equiv (x_0, 0)$. If we join M and B, we get the line of regression of x on y .

If these two regression lines coincide, the correlation is perfect; if they cut at rightangle, the correlation is zero; and if they diverge and intersect each other, the correlation is limited. As the divergence between the two regression lines increases, the correlation decreases.

Exp.(3) A sample of paired variates is given below—

x ; 1 2 3 4 5 6 7

y : 5 13 16 23 33 38 40.

(a) Compute the two regression lines and estimate x for $y=20$.

(b) Represent the data in a scatter diagram and comment on correlation.

(c) Plot the two regression lines on a graph and give an idea of correlation from the angle between the lines.

Sol. (a) The following table shows the calculations of two regression coefficients.

Pair No.	x	$\xi = x - 4$	ξ^2	y	$\eta = y - 25$	η^2	$\xi\eta$
1	1	-3	9	5	-20	400	60
2	2	-2	4	13	-12	144	24
3	3	-1	1	16	-9	81	9
4	4	0	0	23	8	64	0
5	5	1	1	33	8	64	8
6	6	2	4	38	13	169	26
7	7	3	9	40	15	225	45
Totals	—	0	28	—	7	1087	172

Now we have

$$b_{yx} = \frac{\sum \xi \eta - \sum \xi \sum \eta / n}{\sum \xi^2 - (\sum \xi)^2 / n}$$

$$= \frac{172 - 0}{28 - 0} = 6.14$$

$$b_{xy} = \frac{\sum \xi \eta - \sum \xi \sum \eta / n}{\sum \eta^2 - (\sum \eta)^2 / n}$$

$$= \frac{172 - 0}{1087 - (-7)^2 / 7}$$

$$= 0.16$$

Also, $\bar{x} = 4 + \frac{\sum \xi}{n} = 4 + 0 = 4$, and $\bar{y} = 25 + \frac{\sum \eta}{n} = 25 + \frac{(-7)}{7} = 24$.

Thus the regression line of y on x is

$$y - \bar{y} = b_{yx}(x - \bar{x}), \text{ or } y - 24 = 6.14(x - 4), \text{ i.e. } y = 6.14x - 0.56 \dots \text{Ans.}$$

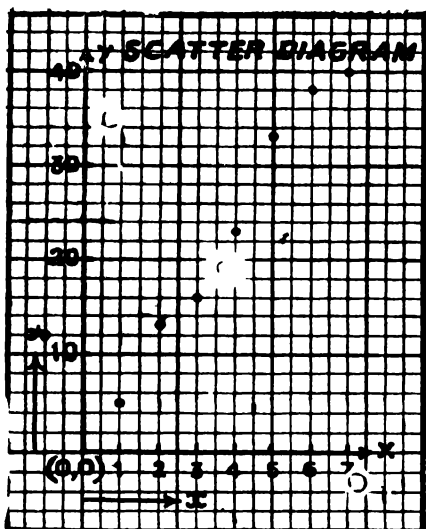
Similarly, the line of regression of x on y is

$$x - \bar{x} = b_{xy}(y - \bar{y}), \text{ or } x - 4 = 0.16(y - 24), \text{ i.e. } x = 0.16y + 0.16 \dots \text{Ans.}$$

and estimate of x for $y=20$ is $x_0 = 0.16 \times 20 + 0.16 = 3.36 \dots \text{Ans.}$

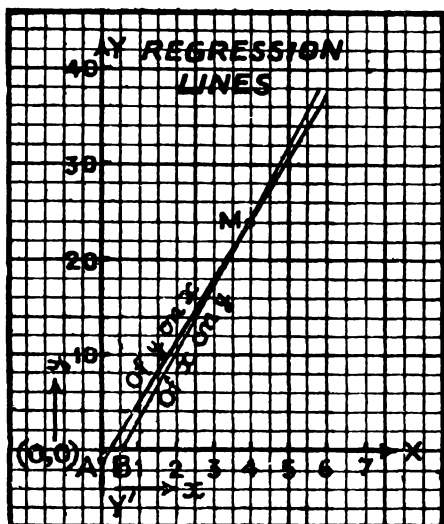
(b) If we measure the variate x along the axis of x , and y along the axis of y , then we get the scatter diagram with seven points [(1,5); (2,13); (3,16); (4,23); (5,33); (6,38); (7,40)] on it. The points seem to be clustered about a line starting from the lower left hand corner

and going to the upper right hand corner (i.e. a line with positive slope), hence the correlation between x and y is limited positive. Since the points are very close to the diagonal with positive slope, so the correlation is of high degree and has a tendency of approaching the perfect value.



(c) We have the two regression lines as
 $y = 6.14x - 0.56 \dots (i)$
 and $x = 0.16y + 0.16 \dots (ii)$,
 from (a).

If we put $x=0$ in (i), we get $A \equiv (0, -0.56)$, and on putting $y=0$ in (ii), we get $B \equiv (0.16, 0)$. The common point of these equations is $M \equiv (\bar{x}, \bar{y}) \equiv (4, 24)$. Thus on joining the points A and M, we get the line of y on x . Similarly, the line of x



only can be drawn by joining the points B and M. Since the angle between these two lines of regression plotted on the graph is an acute angle (ψ) with a very low magnitude, so the correlation is limited positive. As the lines tend to coincide, so the correlation tends to unity.

7.3 Testing the significance of difference between correlation coefficients : It consists of dividing the difference between the two correlation coefficients by the standard error of the difference and then finding the probability of observing this ratio. The comput-

ations of the s.e. and the above mentioned probability depend upon the following two situations—

(a) When the correlation coefficient in the population is zero, i.e. $\rho=0$.

(b) When the correlation coefficient in the population is not zero, i.e. $\rho \neq 0$.

7.3 (a) Testing the significance of an observed correlation coefficient r when $\rho=0$. Let r be the correlation coefficient of a random sample of n pairs drawn from a bivariate normal population with zero correlation coefficient. Then testing the significance of the observed r is the same as testing the significance of difference ($r-0$), or $\rho=0$. Here we usually test the hypothesis (H_0)—that the sample has been taken from a bivariate normal population with zero correlation coefficient, i.e. $\rho=0$. If the hyp. is true, we compute the statistic

$$t = \frac{r}{\sqrt{(1-r^2)/(n-2)}}$$
 which follows a 't' distribution with $(n-2)$ d.f. Here the quantity $\sqrt{[(1-r^2)/(n-2)]}$ is the s.e. of r in a random sample of n . If the absolute value of this statistic i.e. $|t| \geq t_{0.05}(n-2)$, we reject the hyp. at 5% level, otherwise the sample is said to be consistent with the hypothesis.

The test is based upon the following assumptions—

- (1) The sample is a simple random.
- (2) The sample may be large or small.
- (3) The parent population is a bivariate normal.
- (4) The population correlation coefficient is zero.

Note : If r is the observed rank correlation coefficient from a random sample of n pairs, and the hyp. that the population rank correlation coefficient is zero is correct, then the same statistic 't' as stated above is used to test the significance of r . It is otherwise obvious because the rank correlation coefficient can be regarded as the coefficient of correlation between the two variables.

Exp. (4) [a] A sample of 10 villages from Meerut district showed a correlation coefficient of $+0.75$ between 'total cultivable area' and 'area under wheat.' Is this correlation significant? Also find the probable error of r .

[b] Find the least value of r in a random sample of 27 pairs from a bivariate normal population, significant at 5% level.

Sol. [a]. (H_0) : $\rho=0$.

Here we compute the statistic

$$t = \frac{r}{\sqrt{[(1-r^2)/(n-2)]}}$$

$$\text{or } t = \frac{0.75}{\sqrt{[1-(0.75)^2]/(10-2)}} = \frac{0.75}{0.23} = 3.26.$$

Now $|t| = 3.26$ and $t_{.05}(8) = 2.306$.

So $|t| > t_{.05}$ leading to the rejection of hyp. at 5% level.

Also, the probable error of r is given by

$$P. E. (r) = .6745 \times S. E. (r)$$

$$= .6745 \times \sqrt{[(1-r^2)/(n-2)]} = .6745 \times .23 = 0.155.$$

Conclusion : The value of r ($=0.75$) is not significant at 5% level, and its probable error is 0.155.

[b] Let the least significant value of correlation coefficient at 5% level be r . Then we must have $|t| \geq t_{.05}$

$$\text{or } |r|/\sqrt{[(1-r^2)/(n-2)]} \geq t_{.05}(25), \quad \text{i.e. } |5r/\sqrt{(1-r^2)}| \geq 2.06$$

$$\text{or } 25r^2 \geq (2.06)^2 (1-r^2), \quad \text{i.e. } 25r^2 \geq 4.2436 - 4.2436 r^2$$

$$\text{or } 29.2436 r^2 \geq 4.2436, \quad \text{i.e. } r^2 \geq 0.145, \quad \therefore r \geq \pm 0.383.$$

Thus 0.383 is the least value of $|r|$ required..... **Ans.**

Situation (b) when $\rho \neq 0$.

(1) One-sample problem.

$$H_0: \rho = \dots$$

(2) Two-sample problem.

$$H_0: \rho_1 = \rho_2.$$

(3) K-sample problem.

$$H_0: \rho_1 = \rho_2 = \dots = \rho_k.$$

7.3(b) When $\rho \neq 0$: we shall discuss this problem in the following three different situations—

(b-1) One-sample problem :

It is concerned with testing the significance of an observed correlation coefficient when $\rho \neq 0$.

Let r be the correlation coefficient of a large random sample of n pairs drawn from a bivariate normal population with some specified correlation coefficient. Then testing the significance of the observed r is the same as testing the significance of difference $(r-\rho)$, or $\rho = \dots$. Here we usually test the hyp. —that the sample has been taken from a bivariate normal population with specified correlation coefficient, i.e. $\rho = \dots$. If the hyp. is true, we compute the statistic

$$Z = \frac{z-\xi}{1/\sqrt{(n-3)}} \quad \text{which follows a 'standard normal' distribution for large samples.}$$

Here the quantity $1/\sqrt{(n-3)}$ is the s.e. of difference $(z-\xi)$ in a random sample of n . The quantity z is distributed asymptotically normally about the mean ξ with variance

$$1/(n-3), \text{ i.e. } Z \sim AN\left(\xi, \frac{1}{n-3}\right), \text{ where by Fisher's } z\text{-transformation:}$$

$$z = 1.1513 \log_{10}\left(\frac{1+r}{1-r}\right) \quad \text{and} \quad \xi = 1.1513 \log_{10}\left(\frac{1+\rho}{1-\rho}\right). \text{ If } |Z| \geq 1.96,$$

we reject the hyp. at 5% level, otherwise we say that there is no

evidence against the hyp., or the sample is consistent with the hypothesis.

The test is based upon the following assumptions—

- (1) The sample is a simple random.
- (2) The sample is large.
- (3) The parent population is a bivariate normal.
- (4) The population correlation coefficient is specified.

(b-2) Two-sample problem : *It is concerned with testing the significance of difference between two observed correlation coefficients r_1 and r_2 when $\rho \neq 0$.*

Let r_1 and r_2 be the correlation coefficients of two large independent random samples of sizes n_1, n_2 drawn from the same bivariate normal population or from two different populations with the same correlation coefficient. Then testing the significance of difference ($r_1 \sim r_2$) is the same as testing $\rho_1 = \rho_2$, where ρ_1, ρ_2 are the correlation coefficients of the two populations. Here we usually test the hyp.—that the two samples have been taken from two different populations with the same correlation coefficient, i.e. $\rho_1 = \rho_2$. If the hyp. is true, we compute the statistic

$$Z = \frac{z_1 - z_2}{\sqrt{\left(\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}\right)}} \text{ which follows a 'standard normal' }$$

distribution for large samples. Here the quantity $\sqrt{\left(\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}\right)}$ is the s.e. of difference ($z_1 \sim z_2$) in two random and independent samples of sizes n_1, n_2 . The quantities z_1, z_2 are distributed asymptotically normally about the common mean ξ with respective variances $\frac{1}{n_1 - 3}, \frac{1}{n_2 - 3}$; i.e. $Z_1 \sim AN\left(\xi, \frac{1}{n_1 - 3}\right)$ and $Z_2 \sim AN\left(\xi, \frac{1}{n_2 - 3}\right)$

where by Fisher's z -transformations : $z_1 = 1.1513 \log_{10} \left(\frac{1+r_1}{1-r_1}\right)$,
 $z_2 = 1.1513 \log_{10} \left(\frac{1+r_2}{1-r_2}\right)$ and $\xi = 1.1513 \log_{10} \left(\frac{1+\rho}{1-\rho}\right)$.

If $|Z| \geq 1.96$, we reject the hyp. at 5% level, otherwise the two samples are said to be consistent with the hypothesis.

The test is based upon the following assumptions—

- (1) The two samples are simple random and independent.
- (2) The samples are large.
- (3) The parent populations are bivariate normal.
- (4) The correlation coefficients of the two populations are the same.

(b-3) k-sample problem : *It is concerned with testing the*

homogeneity of k (>2) observed correlation coefficients r_1, r_2, \dots, r_k when $\rho \neq 0$.

Let r_1, r_2, \dots, r_k be the correlation coefficients of k large independent random samples of sizes n_1, n_2, \dots, n_k drawn from the same bivariate normal population or from k -different populations with the same correlation coefficient. Then testing the homogeneity of r_1, r_2, \dots, r_k is the same as testing $\rho_1 = \rho_2 = \dots = \rho_k$, where $\rho_1, \rho_2, \dots, \rho_k$ are the correlation coefficients of k -populations. Here we usually test the hyp. - *that the k -samples have been taken from k -different populations with the same correlation coefficient*, i.e. $\rho_1 = \rho_2 = \dots = \rho_k$. If the hyp. is true, we compute the statistic $\chi^2 = \sum (n_i - 3)(z_i - \bar{z})^2$ which follows a ' χ^2 ' distribution with $(k-1)$ d.f. for large samples. Here the quantity z_i ($i=1, 2, \dots, k$) is distributed asymptotically normally about the common mean ξ with variance $\frac{1}{n_i - 3}$, i.e. $z_i \sim$

$AN\left(\xi, \frac{1}{n_i - 3}\right)$, where by Fisher's z -transformations :

$z_i = 1.1513 \log_{10}\left(\frac{1+r_i}{1-r_i}\right)$ and $\xi = 1.1513 \log_{10}\left(\frac{1+\rho}{1-\rho}\right)$. Also, the quantity $\bar{z} = [\sum (n_i - 3)z_i / \sum (n_i - 3)]$, the pooled estimate of ξ , is the weighted mean of z_1, z_2, \dots, z_k with corresponding weights $(n_1 - 3), (n_2 - 3), \dots, (n_k - 3)$ as the reciprocals of their variances in order to estimate ξ with minimum variance. This quantity \bar{z} can also give us the estimate for ρ , using the relation : $\rho = \tanh \bar{z}$.

If $\chi^2 \geq \chi^2_{0.05}$, we reject the hyp. at 5% level, otherwise the samples are said to be consistent with the hypothesis.

The test is based upon the following assumptions —

- (1) The samples are simple random and independent.
- (2) The samples are large.
- (3) The parent populations are bivariate normal.
- (4) The correlation coefficients of the populations are the same.

Exp. (5). [a] A correlation coefficient of 0.73 is obtained from a random sample of 28 pairs. Is this correlation significantly different from 0.5 ?

[b] The correlation coefficients between the temperatures of unhusked rice and breakage percentage calculated from two independent random samples of size 12 and 19 are 0.75 and 0.88 respectively. Do the two estimates differ significantly ?

[c] Three independent samples of 23, 33 and 53 pairs of values give correlation coefficients 0.40, 0.60 and 0.50 respectively. Are these correlation coefficients homogeneous ?

Sol. [a]. (H₀) : $\rho = 0.5$.

Here we compute the statistic

$$Z = \frac{z - \xi}{1/\sqrt{(n-3)}}, \quad \text{where } z = 1.1513 \log_{10} \left(\frac{1 + .73}{1 - .7} \right) = 0.93 \quad \text{and}$$

$$\xi = 1.1513 \log_{10} \left(\frac{1 + .5}{1 - .5} \right) = 0.55.$$

So $|Z| < 1.96$ showing no evidence against the hyp. at 5% level.

Conclusion : The value of r ($= 0.73$) is not significantly different from 0.5 at 5% level of significance.

[b]. (H₀) : $\rho_1 = \rho_2$.

Here we compute the statistic

$$Z = \frac{z_1 - z_2}{\sqrt{\left(\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3} \right)}}, \quad \text{where } z_1 = 1.1513 \log_{10} \left(\frac{1 + .95}{1 - .95} \right) = 1.83$$

$$\text{and } z_2 = 1.1513 \log_{10} \left(\frac{1 + .88}{1 - .88} \right) = 1.38,$$

$$= \frac{1.83 - 1.38}{\sqrt{\left(\frac{1}{9} + \frac{1}{16} \right)}} = 0.45 \times \frac{12}{5} = 1.08.$$

So $|Z| < 1.96$ showing no evidence against the hyp. at 5% level.

Conclusion : The two estimates are not significantly different at 5% level of significance.

[c] (H₀) : $\rho_1 = \rho_2 = \rho_3$.

Here we compute the statistic $\chi^2 = \sum (n_i - 3)(z_i - \bar{z})^2$ from the following table—

Sample No. i	$n_i - 3$	r_i	z_i	$(n_i - 3)z_i$	$(z_i - \bar{z})$	$(z_i - \bar{z})^2$	$(n_i - 3)(z_i - \bar{z})^2$
1	20	.40	.42	8.4	-.15	.0225	.450
2	30	.60	.69	20.7	.12	.0144	.432
3	50	.50	.55	27.5	-.02	.0004	.020
Totals	100	—	—	56.6	—	—	.902 = χ^2

Now we have

$$\chi^2 = \sum (n_i - 3)(z_i - \bar{z})^2 = 0.902, \text{ and } \chi^2_{.05}(2) = 5.991.$$

So $\chi^2 < \chi^2_{.05}$ showing no evidence against the hyp. at 5% level.

Conclusion : The given correlation coefficients do not differ significantly at 5% level of significance, and hence they are homogeneous.

7.4 Testing the significance of an observed regression coefficient b when β is some specified value : Let $b_{y\cdot}$ be the regre

ssion coefficient of y on x in a random sample of n pairs drawn from a bivariate normal population with specified regression coefficient. Then testing the significance of the observed regression coefficient b is the same as testing the significance of difference $(b-\beta)$, or $\beta = \dots$. Here we usually test the hyp.—that the sample has been taken from a bivariate normal population with specified regression coefficient, i.e. $\beta = \dots$. If the hyp. is true, we compute the statistic $t = \frac{b_{yx} - \beta}{\sqrt{[(\sigma_y^2 - b_{yx}^2 \sigma_x^2)/(n-2)\sigma_x^2]}}$ which follows a 't' distribution with $(n-2)$ d.f. Here the quantity $\sqrt{[(\sigma_y^2 - b_{yx}^2 \sigma_x^2)/(n-2)\sigma_x^2]}$ is the s.e. of difference $(b_{yx} - \beta)$ in a random sample of n giving σ_x^2 , σ_y^2 as the variances of x , y respectively. If $|t| \geq t_{.05}(n-2)$, we reject the hyp. at 5% level, otherwise the sample is said to be consistent with the hypothesis.

The regression coefficient of x on y in a random sample of n pairs with corresponding specified value β in the population can be tested similarly by computing the statistic

$$t = \frac{b_{xy} - \beta}{\sqrt{[(\sigma_x^2 - b_{xy}^2 \sigma_y^2)/(n-2)\sigma_y^2]}} \text{ and then comparing } |t| \text{ against } t_{.05}(n-2).$$

The test is based upon the following assumptions—

- (1) The sample is a simple random.
- (2) The sample may be large or small.
- (3) The parent population is a bivariate normal.
- (4) The population regression coefficient is specified.

7.5 Testing the significance of an observed regression function: Let r be the correlation coefficient of a random sample of n pairs drawn from a bivariate normal population with zero correlation coefficient. Then to test the hyp.—that the sample indicates the same degree of association between the two variables as expressed by a regression equation, we compute the statistic

$$F = \frac{r^2}{(1-r^2)/(n-2)} \text{ which follows a 'F' distribution with } (1, n-2) \text{ d.f. If } F \geq F_{.05}(1, n-2), \text{ we reject the hyp. at 5\% level, otherwise the sample is said to be consistent with the hypothesis.}$$

This statistic F , obviously, is the square of the statistic 't' used for testing the significance of r when $\rho=0$, and hence the two tests are equivalent. Therefore, the assumptions for the test are the same as stated for the case when $\rho=0$ in §7.3(a)

Exp. (6) Given that : $n=50$, $S(x-\bar{x})^2=1225$, Calculate b_{yx} and $\bar{x}=30$, $S(y-\bar{y})^2=441$, test its significance.
 $\bar{y}=25$, $S(x-\bar{x})(y-\bar{y})=451$.

Sol : we have $b_{yx} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2} = \frac{451}{1225} = 0.37$.

Ho : $\beta=0$.

Here we compute the statistic

$$t = \frac{b_{yx} - \beta}{\sqrt{[(\sigma_y^2 - b_{yx}^2 \sigma_x^2) / (n-2)\sigma_y^2]}} = \frac{0.37 - 0}{\sqrt{[(441 - (0.37)^2 \times 1225) / 48 \times 1225]}} = \frac{0.370}{0.068} = 5.44.$$

Now $|t| \approx 5.44$ and $t_{0.05}(48) = 2.01$.

So $|t| > t_{0.05}$ leading to the rejection of hyp. at 5% level.

Conclusion : The regression coefficient b_{yx} is 0.37, which is significant at 5% level for $n=50$.

Exp. (7). [a] What is a linear regression ? How the principle of least squares is employed for estimating the constants of regression in a regression equation ?

[b] Give the relation between the regression coefficients and the correlation coefficient. What is the main difference between these two types of coefficients ? Also point out the difference between the correlation and regression theories.

Sol. [a] If the points on a scatter diagram seem to cluster about a straight line, it suggests some linear relationship between the variables, and this straight line is called *the line of regression*. This gives us the average value of the dependent variable corresponding to a given value of the independent variable. If both the variables can take the role of independent variable, we have two regression lines— (1) of y on x , and (2) of x on y . These lines give us the average values of y and x for corresponding given values of x and y respectively. The two lines, in general, are different except in the case when the correlation between x and y is perfect.

Let us now consider $y=a+bx$ be an equation of regression line of y on x in a random sample of n pairs drawn from a bivariate normal population, where a and b are the constants of regression. If $y_e(=a+bx)$ be the estimated value of y for a given value of x , and $nS_y^2=\Sigma(y-y_e)^2=\Sigma(y-a-bx)^2=R$ be the sum of squares of deviations of the observed value y from the estimated value y_e , summation being taken over all pairs of values, then to estimate the constants a and b , the principle of least squares (PLS) is employed. According to this principle, the constants a and b

are so chosen that the residual sum of squares R is minimum.

For this purpose, we partially differentiate R with respect to a and b , and then equate each to zero to get the two normal equations as—

$$\left. \begin{aligned} \frac{\partial R}{\partial a} &= -2\Sigma(y - a - bx) = 0 \\ \frac{\partial R}{\partial b} &= -2\Sigma x(y - a - bx) = 0 \end{aligned} \right\}, \text{ or } \left. \begin{aligned} \Sigma y - na - b\Sigma x &= 0 \\ \Sigma xy - a\Sigma x - b\Sigma x^2 &= 0 \end{aligned} \right\}$$

$$i.e. \quad \left. \begin{aligned} \bar{y} &= a + b\bar{x} \\ \text{cov}(x, y) + \bar{x}\bar{y} &= a\bar{x} + b(\bar{x}^2 + \sigma_x^2) \end{aligned} \right\} \dots\dots (1)$$

$$\dots\dots (2)$$

Multiplying (1) by \bar{x} and subtracting from (2) we get

$b = \text{cov}(x, y) / \sigma_x^2$, and then $a = \bar{y} - \frac{\text{cov}(x, y)}{\sigma_x^2} \bar{x}$. Substituting these estimated values of a and b into the equation $y = a + bx$, the equation to the line of regression of y on x is

$$y - \bar{y} = b_{yx}(x - \bar{x}) \dots\dots (I), \text{ where } b_{yx} = \text{cov}(x, y) / \sigma_x^2.$$

Similarly, the equation to the line of regression of x on y is

$$x - \bar{x} = b_{xy}(y - \bar{y}) \dots\dots (II), \text{ where } b_{xy} = \text{cov}(x, y) / \sigma_y^2.$$

From (I), (II), it is obvious that both the lines pass through the mean point (\bar{x}, \bar{y}) . The coefficients b_{yx} and b_{xy} are called the regression coefficients of y on x , and of x on y respectively.

[b] The correlation coefficient between the two variables x and y is the geometric mean of their two regression coefficients, i.e. $r = \pm \sqrt{(b_{yx} \cdot b_{xy})}$. The sign of r is (+) or (—) according to the sign possessed by either of the regression coefficients.

The main difference between the coefficients of correlation and those of regression may be seen as follows:

(1) The coefficient of correlation is a measure of direction and extent of correlation between the two variables, and it indicates whether the change in one variable tends to bring a change in the other variable in the same or reverse direction. But a coefficient of regression gives the average change in one variable corresponding to a unit change in the other variable.

(2) The coefficient of correlation can never exceed unity while a regression coefficient can.

(3) The coefficient of correlation between any two variables is always symmetrical in the variables but a regression coefficient is rarely so.

(4) The coefficient of correlation is independent of the change of both origin and scale but a regression coefficient is independent of the change of origin only, and not of scale.

The fundamental difference between the theories of correlation and regression may be looked as follows :

(1) In the correlation theory both the variables are assumed as random while in the theory of regression, one of the two variables is treated as independent and the other as dependent.

(2) No correlation implies no regression of the variables but its reverse is not always true.

Exp. (8). [a] Show that ψ , the acute angle between the two lines of regression, is given by : $\tan \psi = (1-r^2)\sigma_y/r(\sigma_x^2 + \sigma_y^2)$. Interpret the case when $r=0, \pm 1$.

[b] Obtain the standard error of estimate of y (or x), and hence or otherwise show that the departure of the value of r^2 from unity is a measure of departure of the relationship between the two variables from linearity.

Sol. [a] If θ_1, θ_2 be the angles which the two regression lines make with the axis of x , then the slopes of the two lines may be given as $\tan \theta_1 = r\sigma_y/\sigma_x = b_{yx}$, and $\tan \theta_2 = \sigma_y/r\sigma_x = 1/b_{xy}$. Thus the acute angle ψ between the two lines of regression is given by

$$\begin{aligned} \tan \psi &= \tan (\theta_2 - \theta_1) = \frac{\tan \theta_2 - \tan \theta_1}{1 + \tan \theta_2 \cdot \tan \theta_1} = \frac{\frac{\sigma_y}{r\sigma_x} - \frac{r\sigma_y}{\sigma_x}}{1 + \frac{\sigma_y}{r\sigma_x} \cdot \frac{r\sigma_y}{\sigma_x}} = \frac{\sigma_y(1-r^2)/r\sigma_x}{r(\sigma_x^2 + \sigma_y^2)/r\sigma_x^2} \quad (\because r^2 \leq 1) \\ &= \frac{1-r^2}{r} \cdot \frac{\sigma_y}{(\sigma_x^2 + \sigma_y^2)} \dots \dots \dots \text{Hence proved.} \end{aligned}$$

If $r=0$, then $\psi=90^\circ$ so that the two lines of regression are perpendicular to each other. In this situation, the estimated value of y (or x) is the same for all values of x (or y). But if $r=\pm 1$, then $\psi=0, 180^\circ$ so that the two lines of regression coincide with each other. **Ans.**

[b] Let $y_e = [\bar{y} + b_{yx}(x - \bar{x})]$ be the estimated value of y corresponding to a given value of x obtained from the regression equation of y on x in a random sample of n drawn from a bivariate normal population. Then the minimum residual sum of squares of y is given by

$$\begin{aligned} nS^2_y &= \Sigma(y - y_e)^2 \\ &= \Sigma[y - \bar{y} - b_{yx}(x - \bar{x})]^2 \\ &= \Sigma(y - \bar{y})^2 - 2b_{yx}\Sigma(x - \bar{x})(y - \bar{y}) + b_{yx}^2\Sigma(x - \bar{x})^2 \\ &= n\sigma_y^2 - 2b_{yx} \cdot nr\sigma_x\sigma_y + b_{yx}^2n\sigma_x^2 \\ &= n\left[\sigma_y^2 - \frac{r\sigma_y}{\sigma_x} \cdot r\sigma_x\sigma_y + \frac{r^2\sigma_y^2}{\sigma_x^2} \cdot \sigma_x^2\right] \\ &= n\sigma_y^2(1-r^2) \end{aligned}$$

$$\therefore S_y^2 = \sigma_y^2(1-r^2), \text{ and } S_y = \sigma_y(1-r^2)^{1/2} \dots\dots\dots (1)$$

This value of S_y , given by (1), is called *the standard error of estimate of y*, or sometimes the root-mean square error of estimate of y . Similarly, the standard error of estimate of x is given by

$$S_x = \sigma_x(1-r^2)^{1/2} \dots\dots\dots (2)$$

Now from (1) or (2), it is obvious that $r^2 \leq 1$, or $-1 \leq r \leq +1$, since the sum of squares of deviations is always positive. Thus if $r = \pm 1$, the sum of squares of deviations from either of the regression lines is zero, and consequently each deviation vanishes showing that all the points lie on both the lines of regression. It means that both the lines then coincide with each other and hence there is a linear functional relation between the two variables x and y under study. Further, as r^2 approaches unity, the residual sum of squares S_x^2 and S_y^2 approach zero so that the points are closer to the regression lines which tend to coincide. Therefore, the departure of the value of r^2 from unity is a measure of departure of the relationship between the two variables from linearity.

Exp. (9). [a] In a partially destroyed laboratory record of an analysis of correlation data, the following results only are legible—

$$\begin{aligned} \text{Variance of } x &= 9, \text{ the regression equ—} & 8x - 10y + 66 &= 0, \\ & & 40x - 18y &= 214. \end{aligned}$$

Find, what were (i) the mean values of x and y ; (ii) the correlation coefficient between x and y ; and (iii) the s.d. of y ?

[b] Find the most likely price in Bombay corresponding to the price of Rs. 70 in Calcutta from the following data—

	Calcutta	Bombay
Average price...	65	67
Standard deviation...	2.5	3.5

Correlation is $+0.8$ between the two prices of the commodities in the two cities.

[c] Given that : $n=50$, $\bar{x}=30$, $\bar{y}=25$, $S(x-\bar{x})^2=1225$, $S(y-\bar{y})^2=441$, and $S(x-\bar{x})(y-\bar{y})=451$. Find the two regression lines and hence or otherwise the coefficient of correlation between x and y .

Sol: [a]. (i) Since we know that the two regression lines intersect at the point (\bar{x}, \bar{y}) , so the means of x and y can be found by solving the equations $8x - 10y + 66 = 0$, and $40x - 18y = 214$. Thus we have : $\left. \begin{aligned} 8x - 10y &= -66 \\ 40x - 18y &= 214 \end{aligned} \right\}, \text{ or } \left. \begin{aligned} 40x - 50y &= -330 \\ 40x - 18y &= 214 \end{aligned} \right\} \text{ giving } x=13, y=17, \text{ i.e. } (x, y) \equiv (\bar{x}, \bar{y}) = (13, 17).$

Hence the mean values of x and y are 13, 17 respectively.....**Ans.**

(ii) In order to find the correlation coefficient r between x and

Also, $r^2 = b_{yx} \times b_{xy} = 0.37 \times 1.02 = 0.3774$

$\therefore r = +0.61 \dots \dots \dots$...Ans.

Aliter : We know that $r = \frac{S(x - \bar{x})(y - \bar{y})}{\sqrt{[S(x - \bar{x})^2 \cdot S(y - \bar{y})^2]}}$

$$= \frac{451}{\sqrt{[1225 \times 441]}} = +0.61 \dots \dots \dots$$
 Ans.

7.6 Multiple and Partial Correlation.

7.6.1 Introduction : In a multivariate population consisting of three or more variables, sometimes we are interested in knowing the relationship between any two variables. In such cases, the different variables may be mutually related by some one or the other phenomenon which will usually be influenced by the other remaining variables of the population. For example, the grain-yields are affected by the sowing-dates, row spacings, depths of ploughing or sowing, levels of irrigation and doses of a fertilizer used. Such type of relationship between any two variables may be studied in the following two ways —

Methods of Study

- (i) Specification method.
- (ii) Elimination method.

(i) Specification method :

Here we consider only those variables of the observed data in which the others have some

specified values. It is usually employed in finding the combined influence of a group of variables upon a variable not the member of the group. The method is useful in the study of *multiple correlation* and *multiple regression*. But it has the disadvantage of restricting the size of the data, and also the results of this can be applied to only those situations wherein the other variables have some specified values.

(ii) Elimination method : Here we eliminate the influence of the other remaining variables on the two variables under study. The method is useful in the study of *partial correlation*. But it has the disadvantage that only the linear effect of the variables can mathematically be eliminated from both the variables under study, and not the entire influence.

7.7 Determination of multiple regression equations : Let us consider a random sample of n sets of corresponding values of the three variables x_1 , x_2 and x_3 drawn from a *trivariate normal population*. If these variables are measured from their respective means \bar{x}_1 , \bar{x}_2 and \bar{x}_3 , i.e. the origin is necessarily at the means, then the quantities so obtained can be represented by x_1 , x_2 and x_3 . Now the multiple regression equation of x_1 on x_2 , x_3 can be given as

$$x_1 = a + b_{12.3}x_2 + b_{13.2}x_3 \dots \dots \dots (1)$$

where the *constants* a and b 's are such as to provide on the average the *best estimate* of x_1 viz. x_{1e} for some specified values of the remaining variables x_2 and x_3 . The quantities a and b 's are called the *constants of regression*.

The constants $b_{12.3}$ and $b_{13.2}$ are known as the *partial regression coefficients* of x_1 on x_2 , and of x_1 on x_3 respectively, which sometimes may be called as the regression coefficients of 1st order. The order is decided by the no. of the subscripts after the point. The subscripts occurring before a point are called the *primary subscripts*, those after the point as *secondary*. Amongst the primary, the first subscript to the b 's denotes the subscript of the dependent variable and the second subscript to that of the x to which it is attached, while the secondary subscripts denote the subscripts of those variables whose effects have been eliminated from those of primary. Thus $b_{12.3}$ is the regression coefficient of 1st order of x_1 on x_2 after the linear effect of the third variable x_3 has been eliminated from both x_1 and x_2 . Or, it may be said as the regression coefficient of $x_{1.3}$ on $x_{2.3}$, where $x_{1.3} (=x_1 - b_{13.3}x_3)$ and $x_{2.3} (=x_2 - b_{23.3}x_3)$ are the *residuals* of 1st order of x_1 on x_3 , and of x_2 on x_3 respectively. The similar meanings are attached with the other coefficient $b_{13.2}$.

The above said constants a and b 's of the equ. (1) can be obtained by employing the principle of least squares. According to this principle, the values of a and b 's are so chosen that the residual sum of squares of x_1 (i.e. the sum of squares of deviations of the observed values x_1 from its corresponding estimated values x_{1e} over all n sets of values of x_2, x_3) is minimum. This residual sum of squares is given by

$$\Sigma(x_1 - x_{1e})^2 = \Sigma(x_1 - a - b_{12.3}x_2 - b_{13.2}x_3)^2 = \Sigma x_{1.23}^2 \quad \dots (2)$$

Thus to obtain the best estimate of x_1 , given by (1), we differentiate (2) partially with respect to $a, b_{12.3}$ and $b_{13.2}$, and then equate each to zero to get the following *three normal equations*—

$\Sigma x_{1.23} = 0, \Sigma x_2 x_{1.23} = 0$, and $\Sigma x_3 x_{1.23} = 0$. Solving* these three equations simultaneously for the three unknown constants $a, b_{12.3}$ and $b_{13.2}$, we get $a = 0$,

$$b_{12.3} = -\frac{\sigma_1 \cdot \Delta_{12}}{\sigma_2 \cdot \Delta_{11}} \quad \text{and} \quad b_{13.2} = -\frac{\sigma_1 \cdot \Delta_{13}}{\sigma_3 \cdot \Delta_{11}} \quad \dots (3) \quad \text{for which}$$

$\Sigma x_{1.23}^2$ is minimum and x_{1e} is the best estimate of x_1 . Here the quantities σ_1, σ_2 and σ_3 are the respective *s.d.s.* of x_1, x_2 and x_3 ; and Δ_{11} ,

* For solution of the normal equations, please see §. 7.12

Δ_{12} , Δ_{13} are the respective *cofactors* of the elements in the 1st row and 1st col; 1st row and 2nd col, 1st row and 3rd col. of the

3×3 determinant $\Delta = \begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{vmatrix}$. This determinant of *total correlation coefficients* (or correlation coefficients of zero orders) is obviously symmetrical in the off-diagonal elements, since $r_{12}=r_{21}$; $r_{23}=r_{32}$ and $r_{31}=r_{13}$ owing to the property of symmetry of the total correlation coefficients. The values of the cofactors and the determinant are given as

$$\Delta_{11} = (-1) \cdot \begin{vmatrix} r_{23} & r_{32} \\ r_{31} & 1 \end{vmatrix} = +(1 - r_{23}^2); \Delta_{12} = (-1) \cdot \begin{vmatrix} r_{21} & r_{23} \\ r_{31} & 1 \end{vmatrix} = -(r_{21} - r_{31}r_{23});$$

$$\Delta_{13} = (-1) \cdot \begin{vmatrix} r_{21} & r_{23} \\ r_{31} & r_{32} \end{vmatrix} = +(r_{21}r_{32} - r_{31}r_{23}); \text{ and } \Delta \equiv \Delta_{11} + r_{12}\Delta_{12} + r_{13}\Delta_{13} = 1 - r_{12}^2 - r_{23}^2 - r_{31}^2 + 2r_{12}r_{23}r_{31}.$$

If we substitute these values of Δ 's into (3), we get the values of the regression coefficients of 1st order in terms of correlation coefficients and s.d.s. of zero order each. Hence on substituting these computed values of the constants a and b 's from (3) into (1), we get the desired equation of the regression plane of x_1 on x_2 , x_3 as

$$x_{1e} = -\frac{\sigma_1}{\sigma_2} \cdot \frac{\Delta_{12}}{\Delta_{11}} x_2 - \frac{\sigma_1}{\sigma_3} \cdot \frac{\Delta_{13}}{\Delta_{11}} x_3, \text{ i.e. } \frac{x_{1e}}{\sigma_1} \Delta_{11} + \frac{x_2}{\sigma_2} \Delta_{12} + \frac{x_3}{\sigma_3} \Delta_{13} = 0 \quad (4)$$

If the origin is not necessarily at the the means, the above equation of the regression plane will be as

$$\frac{x_{1e} - \bar{x}_1}{\sigma_1} \Delta_{11} + \frac{x_2 - \bar{x}_2}{\sigma_2} \Delta_{12} + \frac{x_3 - \bar{x}_3}{\sigma_3} \Delta_{13} = 0 \dots \dots \dots (4')$$

The multiple regression equations of x_2 on x_1 , x_3 ; and of x_3 on x_1 , x_2 can be obtained similarly. Generalizing this trivariate* case to that of a p -variates, we can show that the equ. of the regression plane of x_1 on x_2, x_3, \dots, x_p will be as

$$\frac{x_{1e}}{\sigma_1} \Delta_{11} + \frac{x_2}{\sigma_2} \Delta_{12} + \dots + \frac{x_p}{\sigma_p} \Delta_{1p} = 0 \dots (5) \text{ (if origin is at the means)}$$

$$\text{or } \frac{x_{1e} - \bar{x}_1}{\sigma_1} \Delta_{11} + \frac{x_2 - \bar{x}_2}{\sigma_2} \Delta_{12} + \dots + \frac{x_p - \bar{x}_p}{\sigma_p} \Delta_{1p} = 0 \dots \dots \dots (5')$$

(if origin is not at the means)

where \bar{x} 's, σ 's are the means, s.d.s. of the corresponding variables x 's; and Δ 's are the *minors* of the corresponding elements in the $p \times p$ determinate Δ of total correlation coefficients.

* The theory of trivariate case was developed by prof. K. Pearson (1896) and a year later its generalization was given by Yule.

7.8 Properties of the residuals : The residuals of any order have the following three main properties—

1st : The sum of the products of the corresponding values of a variate and a residual is always zero provided the subscript of the variate occurs among the secondary subscripts of the residual. For example, $\Sigma x_{1 \cdot 2 \cdot 13} = 0 = \Sigma x_3 x_{2 \cdot 13}$, and $\Sigma x_2 x_{3 \cdot 12} = 0 = \Sigma x_1 x_{3 \cdot 12}$ etc. In general, we have $\Sigma x_i x_{1 \cdot 23 \dots p} = 0$ etc. for $i=2, 3 \dots p$.

2nd: The sum of the products of any two residuals remains unchanged provided we remove from one residual any or all of the secondary subscripts which are common to both. For example,

$$\Sigma x_{1 \cdot 23} x_{1 \cdot 2} = \Sigma x_{1 \cdot 23} (x_1 - b_{12 \cdot 3} x_2) = \Sigma x_{1 \cdot 23} x_1, \text{ and}$$

$$\Sigma x_{1 \cdot 23} x_{1 \cdot 23} = \Sigma x_{1 \cdot 23} (x_1 - b_{12 \cdot 3} x_2 - b_{13 \cdot 2} x_3) = \Sigma x_{1 \cdot 23} x_1 \text{ etc.}$$

In general, we have

$$\Sigma x_{1 \cdot 34 \dots p} x_{2 \cdot 34 \dots p} = \Sigma x_{1 \cdot 34 \dots p} x_2 = \Sigma x_1 x_{2 \cdot 34 \dots p} \text{ etc.}$$

3rd: The sum of the products of any two residuals is always zero provided all the subscripts of a residual occur among the secondary subscripts of the other. For example,

$$\Sigma x_{1 \cdot 23} x_{2 \cdot 3} = \Sigma x_{1 \cdot 23} (x_2 - b_{23} x_3) = 0, \text{ and}$$

$$\Sigma x_{1 \cdot 23} x_{3 \cdot 2} = \Sigma x_{1 \cdot 23} (x_3 - b_{32} x_2) = 0 \text{ etc.}$$

In general, we have $\Sigma x_{1 \cdot 234 \dots p} x_{2 \cdot 34 \dots p} = 0$ etc.

7.9 Determination of residual variances : Let us consider a random sample of n sets of values (x_1, x_2, x_3) drawn from a trivariate normal population. If the variables (x_1, x_2, x_3) are measured from their respective means $(\bar{x}_1, \bar{x}_2, \bar{x}_3)$, then the residual variances of these variables can be represented in terms of their s.d.s. and correlation coefficients of zero order each. For example, the variance of the second order residual of x_1 on x_2, x_3 can be found as follows—

$$\text{var } (x_{1 \cdot 23}) = \frac{1}{n} \Sigma x_{1 \cdot 23}^2 = \frac{1}{n} \Sigma x_{1 \cdot 23} x_{1 \cdot 23} = \frac{1}{n} \Sigma x_1 x_{1 \cdot 23} \text{ (by 2nd property)}$$

$$\text{or } \sigma_{1 \cdot 23}^2 = \frac{1}{n} \Sigma x_1 (x_1 - b_{12 \cdot 3} x_2 - b_{13 \cdot 2} x_3)$$

$$= \frac{1}{n} [\Sigma x_1^2 - b_{12 \cdot 3} \Sigma x_1 x_2 - b_{13 \cdot 2} \Sigma x_1 x_3]$$

$$= \frac{1}{n} [n\sigma_1^2 - b_{12 \cdot 3} n\sigma_1\sigma_2 r_{12} - b_{13 \cdot 2} n\sigma_1\sigma_3 r_{13}]$$

$$= \sigma_1^2 - b_{12 \cdot 3} \sigma_1 \sigma_2 r_{12} - b_{13 \cdot 2} \sigma_1 \sigma_3 r_{13}$$

$$\therefore \sigma_{1 \cdot 23}^2 = \sigma_1^2 \frac{\Delta}{\Delta_{11}} \dots \dots \dots (1)$$

where the symbols have their usual meanings, as stated for relation (3) in § 7.7.

Similarly, it can be shown that the $\text{var}(x_{2.13}) \text{ i.e. } \sigma_{2.13}^2 = \sigma_2^2 \frac{\Delta}{\Delta_{22}}$, and $\text{var}(x_{3.12}) \text{ i.e. } \sigma_{3.12}^2 = \sigma_3^2 \frac{\Delta}{\Delta_{33}}$.

Here it may also be noted that $\text{var}(x_{2.1}) \text{ i.e. } \sigma_{2.1}^2 = \sigma_2^2 \frac{\Delta}{\Delta_{22}} = \sigma_2^2 (1 - r_{12}^2)$

, where $\Delta = \begin{vmatrix} 1 & r_{12} \\ r_{21} & 1 \end{vmatrix} 2 \times 2$, and $\Delta_{22} = 1$. Thus evidently, it is a

residual variance of 1st order only and so may be compared with the expression $S_y^2 = \sigma_y^2 (1 - r^2)$ of (1) in example (8); provided we treat x_2 as y , and x_1 as x . It means that $\sigma_{2.1}$, the s.d. of 1st order of x_2 on x_1 , may be termed as the *s.e. of estimate of x_2 on x_1* when the regression of x_2 (or x_1) on x_1 (or x_2) is linear.

Generalizing this trivariate case to that of a p -variates, we can show that $\text{var}(x_{1.23 \dots p}) \text{ i.e. } \sigma_{1.23 \dots p}^2 = \sigma_1^2 \frac{\Delta}{\Delta_{11}}$, where Δ_{11} is the minor of the element in the 1st row and 1st col. of Δ , the $p \times p$ determinant of total correlation coefficients.

7.10 Determination of multiple correlation coefficients :

A multiple correlation coefficient of a variable on the group of remaining variables under study can be defined as the amount of total correlation between the observed and estimated values of the variable on a regression plane.

Let us now consider a random sample of n sets of corresponding values of the three variables x_1 , x_2 and x_3 drawn from a trivariate normal population. If the variables are measured from their respective means and the quantities thus obtained are represented by x_1 , x_2 and x_3 , then the multiple correlation coefficients of these variables on the groups of remaining two variables can be represented either in terms of residual variances or total correlation coefficients. For example the multiple correlation coefficient of x_1 on x_2 , x_3 , by definition, can be seen as the total correlation coefficient between x_1 and x_{1e} ($= b_{12.3}x_2 + b_{13.2}x_3 = x_1 - x_{1.23}$), where x_{1e} is the estimated value of x_1 for some specified values of x_2 and x_3 on the regression plane. It is generally denoted by the symbol $R_{1(23)}$. Thus by definition, we have

$$\begin{aligned} R_{1(23)} &= \frac{\text{cov}(x_1, x_{1e})}{\sigma_{x_1} \cdot \sigma_{x_{1e}}} \\ &= \frac{\sum x_1 x_{1e}}{\sqrt{[\sum x_1^2 \cdot \sum x_{1e}^2]}} = \frac{\sum x_1 (x_1 - x_{1.23})}{\sqrt{[\sum x_1^2 \cdot \sum (x_1 - x_{1.23})^2]}} \end{aligned}$$

$$\begin{aligned}
 \text{or } R_{1(23)} &= \frac{\Sigma x_1^2 - \Sigma x_1 x_{1.23}}{\sqrt{[\Sigma x_1^2 (\Sigma x_1^2 - 2 \Sigma x_1 x_{1.23} + \Sigma x_{1.23}^2)]}} \\
 &= \frac{(\Sigma x_1^2 - \Sigma x_{1.23}^2)}{\sqrt{[\Sigma x_1^2 (\Sigma x_1^2 - \Sigma x_{1.23}^2)]}} \dots\dots\dots (\text{by 2nd property}) \\
 &= \frac{\sqrt{(\Sigma x_1^2 - \Sigma x_{1.23}^2)}}{\sqrt{(\Sigma x_1^2)}} = \left(1 - \frac{\Sigma x_{1.23}^2}{\Sigma x_1^2}\right)^{1/2} \\
 \therefore R_{1(23)} &= \left(1 - \frac{\sigma_{1.23}^2}{\sigma_1^2}\right)^{1/2}; \text{ or } R_{1(23)}^2 = 1 - \frac{\sigma_{1.23}^2}{\sigma_1^2}; \\
 &\quad \text{i.e. } 1 - R_{1(23)}^2 = \sigma_{1.23}^2 / \sigma_1^2 \dots\dots\dots (1)
 \end{aligned}$$

If in the above relation (1) we substitute the value of residual variance $\sigma_{1.23}^2 (= \sigma_1^2 \Delta / \Delta_{11})$, we can have the value of the multiple correlation coefficient $R_{1(23)}$ in terms of total correlation coefficients as

$$\begin{aligned}
 R_{1(23)} &= \left(1 - \frac{\Delta}{\Delta_{11}}\right)^{1/2}; \text{ or } R_{1(23)}^2 = 1 - \frac{\Delta}{\Delta_{11}}; \text{ i.e. } 1 - R_{1(23)}^2 \\
 &= \Delta / \Delta_{11} \dots\dots (2), \text{ where the symbols have their usual meanings.}
 \end{aligned}$$

Here it may also be noted that $0 \leq R_{1(23)} \leq 1$ always, since the term Σx_1^2 cannot be $(-)$ ve. Further, if $R_{1(23)} = 0$, x_1 is uncorrelated with any of the other variables x_2 and x_3 . But if $R_{1(23)} = 1$, then we have $\sigma_{1.23}^2 = 0$, i.e. all the residuals $x_{1.23}$ are zero, the observed and estimated values of x_1 coincide and hence the observed x_1 is a linear function of x_2 and x_3 .

The multiple correlation coefficients of x_2 on x_3, x_1 ; and of x_3 on x_1, x_2 can be obtained similarly as

$$\begin{aligned}
 R_{2(31)} &= \left(1 - \frac{\sigma_{2.13}^2}{\sigma_2^2}\right)^{1/2} \text{ or } \left(1 - \frac{\Delta}{\Delta_{22}}\right)^{1/2}; \text{ and } R_{3(12)} = \left(1 - \frac{\sigma_{3.12}^2}{\sigma_3^2}\right)^{1/2} \\
 \text{or } &\left(1 - \frac{\Delta}{\Delta_{33}}\right)^{1/2}. \text{ Generalizing this trivariate case to that of a } p\text{-variates, we can show that the multiple correlation coefficient of } x_1 \text{ on }
 \end{aligned}$$

x_2, x_3, \dots, x_p is $R_{1(23 \dots p)} = \left(1 - \frac{\sigma_{1.23 \dots p}^2}{\sigma_1^2}\right)^{1/2} \text{ or } \left(1 - \frac{\Delta}{\Delta_{11}}\right)^{1/2}$, where Δ_{11} is the minor of the element in the 1st row and 1st col. of Δ , the $p \times p$ determinant of total correlation coefficients.

7.10.1 Testing the significance of an observed multiple correlation coefficient R when $\bar{R} = 0$: Let R be the multiple correlation coefficient of a variable on the group of other remaining p -variables in a random sample of n sets of values drawn from a $(p+1)$ -variate normal population with zero multiple correlation coefficient. Then testing the significance of the observed R is the same as testing the significance of difference ($R \sim 0$), or $\bar{R} = 0$. Here we usually test the hypothesis—that the sample has been taken from

a $(p+1)$ -variate normal population with zero multiple correlation coefficient, i.e. $\bar{R}=0$. If the hyp. is true, we compute the statistic (due to Fisher)

$$F = \frac{R^2}{1-R^2} \cdot \frac{n-p-1}{p} \text{ which follows an 'F' distribution with } (p, n-p-1) \text{ d.f.}$$

If $F \geq F_{.05}(p, n-p-1)$, we reject the hyp. at 5% level, otherwise the sample is said to be consistent with the hypothesis.

The test is based on the following assumptions—

- (1) The sample is a simple random.
- (2) The sample may be large or small.
- (3) The parent population is a $(p+1)$ -variate normal.
- (4) The population multiple correlation coefficient is zero
- (5) The quantities R^2 and $(1-R^2)$ are distributed independently like chisquares with p and $(n-p-1)$ d.f. respectively.

7.11 Determination of partial correlation coefficients:

A partial correlation coefficient between any two variables can be defined as the amount of total correlation between these two variables provided the linear effect of the group of remaining variable on a regression plane has been eliminated from both of these variables under study.

Let us now consider a random sample of n sets of corresponding values of the three variables x_1 , x_2 and x_3 drawn from a trivariate normal population. If the variables are measured from their respective means and the quantities thus obtained are denoted by x_1 , x_2 and x_3 , then the partial correlation coefficients between any two variables, when the linear effects of the remaining third variables have been eliminated from both of these variables, can be represented in terms of correlation coefficients of lower order, or the total correlation coefficients. For example, the partial correlation coefficient between x_1 and x_2 , when the linear effect of x_3 has been eliminated from both of x_1 , x_2 , can be seen as the total correlation coefficient between $x_{1.3} (=x_1 - b_{13}x_3)$ and $x_{2.3} (=x_2 - b_{23}x_3)$. It is generally denoted by the symbol $r_{12.3}$, which is obviously a correlation coefficient of 1st order. Thus by definition, we have

$$\begin{aligned} r_{12.3} &= \frac{\text{COV}(x_{1.3}, x_{2.3})}{\sigma_{1.3} \times \sigma_{2.3}} \\ &= \left[\frac{\text{COV}(x_1, x_2)}{\sigma_{1.3} \times \sigma_{2.3}} \times \frac{\text{COV}(x_1, x_2)}{\sigma_{1.3} \times \sigma_{2.3}} \right]^{1/2} \\ &= \left[\frac{\text{COV}(x_1, x_2)}{\sigma_{1.3}^2} \times \frac{\text{COV}(x_2, x_1)}{\sigma_{2.3}^2} \right]^{1/2} = [b_{12.3} \times b_{21.3}]^{1/2} \\ &= \left[\frac{-\sigma_1}{\sigma_2} \cdot \frac{\Delta_{12}}{\Delta_{11}} \times \frac{-\sigma_2}{\sigma_1} \cdot \frac{\Delta_{21}}{\Delta_{22}} \right]^{1/2} = \pm \frac{\Delta_{12}}{\sqrt{(\Delta_{11}\Delta_{22})}} \quad (\because \Delta_{12} \equiv \Delta_{21}) \end{aligned}$$

$$\text{or } r_{12.3} = \frac{-\Delta_{12}}{\sqrt{(\Delta_{11}\Delta_{22})}} \dots\dots\dots(1)$$

$$\therefore r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{[(1-r_{13}^2)(1-r_{23}^2)]}} \dots\dots\dots(2)$$

In relation (1), out of \pm signs we have detained only the $(-)$ ve sign for $r_{12.3}$ because $r_{12.3}$ bears the same sign as possessed by $b_{12.3}$, $b_{21.3}$ and Δ_{12} (or Δ_{21}). The sign of $b_{12.3}$ (or $b_{21.3}$) is $(-)$ ve, as attached with Δ_{12} (or Δ_{21}), since Δ_{11} , Δ_{22} are always $(+)$ ve.

Here it may also be noted that $-1 \leq r_{12.3} \leq +1$ always. Further, if $r_{12.3} = 0$, x_1 is uncorrelated with any of the other variables x_2 and x_3 (or x_2 is uncorrelated with any of the other variables x_3 and x_1). It means that $r_{12.3}$ will not be zero unless $r_{12} = 0$ and at least one of r_{13} , r_{23} is zero. But if $r_{12.3} = 1$, the three regression planes will coincide with each other, and hence the *necessary and sufficient condition* for the coincidence of x_1 , x_2 and x_3 is that $r_{12}^2 + r_{23}^2 + r_{31}^2 - 2r_{12}r_{23}r_{31} = 1$. We also have $r_{12.3} = r_{21.3}$, since a partial correlation coefficient is always symmetrical for the interchange of its primary subscripts provided the secondary subscripts remain the same.

The partial correlation coefficients between x_2 and x_3 after eliminating the effect of x_1 , and between x_3 and x_1 after eliminating the effect of x_2 , can be obtained similarly as

$$r_{23.1} = \frac{r_{23} - r_{21}r_{31}}{\sqrt{[(1-r_{21}^2)(1-r_{31}^2)]}}, \text{ and } r_{31.2} = \frac{r_{31} - r_{32}r_{12}}{\sqrt{[(1-r_{32}^2)(1-r_{12}^2)]}}.$$

If we further consider a similar case of four variables x_1 , x_2 , x_3 and x_4 , then the partial correlation coefficient between x_1 and x_2 after eliminating the effects of x_3 and x_4 can be seen as the amount of total correlation between $x_{1.34}$ and $x_{2.34}$ as

$$r_{12.34} = \frac{r_{12.3} - r_{13.4}r_{23.4}}{\sqrt{[(1-r_{13.4}^2)(1-r_{23.4}^2)]}}, \text{ which is a correlation coefficient of}$$

2nd order expressed in terms of 1st order correlation coefficients.

Generalizing it to a case of p -variables, we can show that the partial correlation coefficient between x_1 and x_2 , after the linear effects of the remaining $(p-2)$ variables x_3, x_4, \dots, x_p have been eliminated from both x_1 and x_2 , can be given as

$$r_{12.34\dots p} = \frac{r_{12.34\dots(p-1)} - r_{13.45\dots p}r_{23.45\dots p}}{\sqrt{[(1-r_{13.45\dots p}^2)(1-r_{23.45\dots p}^2)]}}, \text{ which is a correlation coefficient of } (p-2) \text{ th order expressed in terms of } (p-3) \text{ th order correlation coefficients. It can also be given by the relation}$$

$$r_{12.q} = \frac{\text{COV}(x_{1.q}, x_{2.q})}{\sigma_{1.q} \times \sigma_{2.q}} = (b_{12.q} \times b_{21.q})^{1/2}, \text{ where } q \text{ stands for the group of suffixes } 3, 4, \dots, p.$$

7.11.1 Testing the significance of an observed partial correlation coefficient r when $\rho=0$: Let r be the partial correlation coefficient between any two variables, after eliminating the effects of the remaining k -variables, in a random sample of n sets of values drawn from a $(k+2)$ -variate normal population with zero (corresponding) partial correlation coefficient. Or,

Let r be the k th order correlation coefficient in a random sample of n drawn from a $(k+2)$ -variate normal population with corresponding correlation coefficient as zero. Then testing the significance of the observed r is the same as testing the significance of difference ($r \neq 0$), or $\rho=0$. Here we usually test the hyp.—*that the sample has been taken from a $(k+2)$ -variate normal population with zero correlation coefficient, i.e. $\rho=0$* . If the hyp. is true, we compute the statistic (due to Fisher)

$$t = \frac{r}{\sqrt{[(1-r^2)/(n-k-2)]}}$$
 which follows a 't' distribution with $(n-k-2)$ d.f. Here the quantity $\sqrt{[(1-r^2)/(n-k-2)]}$ is the s.e. of r in a random sample of n . If the absolute value of this statistic i.e. $|t| \geq t_{.05}(n-k-2)$, we reject the hyp. at 5% level, otherwise the sample is said to be consistent with the hypothesis.

The test is based on the following assumptions—

- (1) The sample is a simple random.
- (2) The sample may be large or small.
- (3) The parent population is a $(k+2)$ -variate normal.
- (4) The population partial correlation coefficient is zero.

Note (1) : If $\rho=0$, then the statistic required to test the significance of a k th order correlation coefficient in a random sample of n from a $(k+2)$ -variate normal population is analogous to that of a zero order correlation coefficient from a bivariate normal population with sample size n reduced by k . Thus a test of significance for a partial correlation coefficient with k -secondary subscripts can easily be obtained from that of discussed in § 7.3(a) simply by replacing the quantity $(n-2)$ by $(n-k-2)$.

Note (2) : If $\rho \neq 0$, then the statistics discussed in § 7.3(b-1,2,3,) also hold equally for partial correlation coefficients with the mere change that the sample sizes are further reduced by the nos. of the secondary subscripts of the partial correlation coefficients under study. Thus the tests of significance for the correlation coefficients of order k can easily be obtained from those of discussed in § 7.3(b-1,2,3) simply by replacing the quantities $(n-3)$, (n_1-3) by $(n-k-3)$, (n_1-k-3) respectively.

7.12 Determination of partial regression coefficients : Let us consider a random sample of n sets of corresponding values of the three variables x_1 , x_2 and x_3 drawn from a trivariate normal population. If these variables are measured from their respective means \bar{x}_1, \bar{x}_2 and \bar{x}_3 , and the quantities thus obtained are represented by x_1, x_2 and x_3 , then a multiple regression equation of x_1 on x_2, x_3 can be given as

$$x_{1e} = a + b_{12.3}x_2 + b_{13.2}x_3 \dots \dots \dots (1)$$

where a and b 's are the constants of regression and x_{1e} is the best estimate of x_1 for some given values of x_2 and x_3 . The constants $b_{12.3}$ and $b_{13.2}$ are the *partial regression coefficients* of the above said regression equation. The coefficients of regression along with the constant a need be determined in a way so as to give on the average the best estimate of x_1 corresponding to any assigned values of x_2 and x_3 . These coefficients can be determined in terms of

(a) *s.d.s. and correlation coefficients of zero orders, and*

(b) *s.d.s. and correlation coefficients of higher than zero orders, i.e. partial s.d.s. and partial correlation coefficients.*

(a) b 's in terms of σ , and r , of zero orders :

In order to determine the regression coefficients $b_{12.3}$ and $b_{13.2}$ in terms of $s.d.s.$ and correlation coefficients of zero orders, we have to find the regression constants a and b 's such that

$$R = \Sigma(x_1 - x_{1e})^2 = \Sigma(x_1 - a - b_{12.3}x_2 - b_{13.2}x_3)^2 = \Sigma x_1^2 - 2a \Sigma x_1 - 2b_{12.3} \Sigma x_1 x_2 - 2b_{13.2} \Sigma x_1 x_3 + a^2 n + b_{12.3}^2 \Sigma x_2^2 + b_{13.2}^2 \Sigma x_3^2 + 2ab_{12.3} \Sigma x_2 + 2ab_{13.2} \Sigma x_3 + 2b_{12.3}b_{13.2} \Sigma x_2 x_3 \dots \dots \dots (2)$$

is a minimum so that x_{1e} is the best estimate of x_1 for some given values of x_2 and x_3 . This object can be attained by employing the principle of least squares. According to this principle, we partially differentiate the residual sum of squares of x_1 , given by (2), w.r.t. a and b 's and then equate each to zero in order to have the following three normal equations—

$$(i) \quad \frac{\partial R}{\partial a} = -2 \Sigma(x_1 - a - b_{12.3}x_2 - b_{13.2}x_3) = 0$$

$$(ii) \quad \frac{\partial R}{\partial b_{12.3}} = -2 \Sigma x_2(x_1 - a - b_{12.3}x_2 - b_{13.2}x_3) = 0$$

$$(iii) \quad \frac{\partial R}{\partial b_{13.2}} = -2 \Sigma x_3(x_1 - a - b_{12.3}x_2 - b_{13.2}x_3) = 0$$

or

$$\left. \begin{aligned} \Sigma(x_1 - a - b_{12.3}x_2 - b_{13.2}x_3) &= 0 = \Sigma x_1 - na - b_{12.3} \Sigma x_2 - b_{13.2} \Sigma x_3 \\ \Sigma x_2(x_1 - a - b_{12.3}x_2 - b_{13.2}x_3) &= 0 = \Sigma x_2 x_1 - a \Sigma x_2 - b_{12.3} \Sigma x_2^2 - b_{13.2} \Sigma x_2 x_3 \\ \Sigma x_3(x_1 - a - b_{12.3}x_2 - b_{13.2}x_3) &= 0 = \Sigma x_3 x_1 - a \Sigma x_3 - b_{12.3} \Sigma x_3 x_2 - b_{13.2} \Sigma x_3^2 \end{aligned} \right\}$$

Considering the summation over all n sets of values, we have

$$\left. \begin{aligned} \Sigma x_1 - na - b_{12.3} \Sigma x_2 - b_{13.2} \Sigma x_3 &= 0 \\ \Sigma x_2 x_1 - a \Sigma x_2 - b_{12.3} \Sigma x_2^2 - b_{13.2} \Sigma x_2 x_3 &= 0 \\ \Sigma x_3 x_1 - a \Sigma x_3 - b_{12.3} \Sigma x_3 x_2 - b_{13.2} \Sigma x_3^2 &= 0 \end{aligned} \right\}$$

or

$$\left. \begin{aligned} a=0, \text{ since } \Sigma x_1 = \Sigma x_2 = \Sigma x_3 = 0, \text{ and} \\ nr_{12}\sigma_2\sigma_1 - 0 - b_{12.3}n\sigma_2^2 - b_{13.2}nr_{23}\sigma_2\sigma_3 = 0 \\ nr_{31}\sigma_3\sigma_1 - 0 - b_{12.3}nr_{32}\sigma_3\sigma_2 - b_{13.2}n\sigma_3^2 = 0 \end{aligned} \right\}$$

Thus we have $a=0$ from (i), and $b_{12.3}$, $b_{13.2}$ can be obtained by simplifying the last two equations—

$$\left. \begin{aligned} r_{12}\sigma_1 - b_{12.3}\sigma_2 - b_{13.2}r_{23}\sigma_3 = 0 \\ r_{31}\sigma_1 - b_{12.3}r_{32}\sigma_2 - b_{13.2}\sigma_3 = 0 \end{aligned} \right\}, \text{ or } \left. \begin{aligned} -r_{12}\sigma_1 + b_{12.3}\sigma_2 + b_{13.2}r_{23}\sigma_3 = 0 \\ -r_{31}\sigma_1 + b_{12.3}r_{32}\sigma_2 + b_{13.2}\sigma_3 = 0 \end{aligned} \right\}$$

Now we get

$$b_{12.3} = \frac{\begin{vmatrix} -r_{12}\sigma_1 & r_{23}\sigma_3 \\ -r_{31}\sigma_1 & \sigma_3 \end{vmatrix} \div \begin{vmatrix} \sigma_2 & r_{23}\sigma_3 \\ r_{32}\sigma_2 & \sigma_3 \end{vmatrix}}{\begin{vmatrix} r_{12} & r_{23} \\ r_{31} & 1 \end{vmatrix} \div \sigma_2\sigma_3} = \frac{\begin{vmatrix} 1 & r_{23} \\ r_{32} & 1 \end{vmatrix}}{\sigma_2 \Delta_{11}} = \frac{-\sigma_1 \cdot \Delta_{12}}{\sigma_2 \Delta_{11}},$$

and

$$b_{13.2} = \frac{\begin{vmatrix} -r_{12}\sigma_1 & \sigma_2 \\ -r_{31}\sigma_1 & r_{32}\sigma_2 \end{vmatrix} \div \begin{vmatrix} \sigma_2 & r_{23}\sigma_3 \\ r_{32}\sigma_2 & \sigma_3 \end{vmatrix}}{\begin{vmatrix} r_{12} & 1 \\ r_{31} & r_{32} \end{vmatrix} \div \sigma_2\sigma_3} = \frac{\begin{vmatrix} 1 & r_{23} \\ r_{32} & 1 \end{vmatrix}}{\sigma_3 \Delta_{11}} = \frac{-\sigma_1 \Delta_{13}}{\sigma_3 \Delta_{11}},$$

where Δ_{11} , Δ_{12} , Δ_{13} are the respective cofactors of the elements in the 1st row and 1st col., 1st row and 2nd col.; 1st row and 3rd col. of the 3×3 determinant Δ of total correlation coefficients. Thus the values of $b_{12.3}$ and $b_{13.2}$ are expressible in terms of σ , and r , of zero orders. Similarly,

it can be shown that $b_{23.1} = \frac{-\sigma_2}{\sigma_3} \cdot \frac{\Delta_{23}}{\Delta_{22}}$, and $b_{21.3} = \frac{-\sigma_2}{\sigma_1} \cdot \frac{\Delta_{21}}{\Delta_{22}}$;

etc. Generalizing this trivariate case to that of a p -variates, we

can show that $b_{22.34 \dots p} = \frac{-\sigma_1}{\sigma_2} \cdot \frac{\Delta_{12}}{\Delta_{11}}$, where Δ_{11} , Δ_{12} are the respective

minors of the elements in the 1st row and 1st col.; 1st row and 2nd col. of the $p \times p$ determinant Δ of total correlation coefficients.

(b) b 's in terms of σ , and r , of higher than zero orders :

In order to express the regression coefficients $b_{12.3}$ and $b_{13.2}$ in terms of s.d.s. and correlation coefficients of higher than zero orders *i.e.* in partial values of σ , and r , we have to consider the relations $\Sigma x_2 \cdot x_{1.2} = 0 \dots (1)$, and $\Sigma x_3 \cdot x_{1.3} \dots (2)$ (by 3rd property). From (1), we have

$$\Sigma x_2 \cdot (x_1 - b_{12.3}x_2 - b_{13.2}x_3) = 0$$

$$\text{or } \Sigma x_1 x_2 - b_{12.3} \Sigma x_2 x_2 - b_{13.2} \Sigma x_3 x_2 = 0$$

$$\text{or } \Sigma x_1 \cdot x_2 - b_{12.3} \Sigma x_2 \cdot x_2 - 0 = 0, \quad (\text{by 2nd, 1st properties})$$

$$\text{or } b_{12.3} = \frac{\Sigma x_1 x_2}{\Sigma x_2 x_2}, \text{ or } \frac{\text{cov}(x_1, x_2)}{\text{var}(x_2)},$$

$$\therefore b_{12.3} = r_{12.3} \sigma_1 / \sigma_2 \dots \dots \dots (i)$$

Similarly from (2), we have

$$b_{13.2} = \frac{\Sigma x_{1.2} x_{3.2}}{\Sigma x_{1.2}^2}, \text{ or } \frac{\text{cov}(x_{1.2}, x_{3.2})}{\text{var}(x_{3.2})}$$

$$\therefore b_{13.2} = r_{13.2} \sigma_{1.2} / \sigma_{3.2} \dots \dots \dots (ii)$$

Thus from (i), (ii) we see that the two regression coefficients of order one each can be expressed in terms of σ , and r , each of order one, i.e. in partial values of σ , and r .

$$\text{Similarly, it can be shown that } b_{23.1} = \frac{\Sigma x_{2.1} x_{3.1}}{\Sigma x_{2.1}^2}, \text{ or } \frac{\text{cov}(x_{2.1}, x_{3.1})}{\text{var}(x_{3.1})}$$

i.e. $r_{23.1} \sigma_{2.1} / \sigma_{3.1}$; and $b_{21.3} = \frac{\Sigma x_{2.3} x_{1.3}}{\Sigma x_{2.3}^2}, \text{ or } \frac{\text{cov}(x_{2.3}, x_{1.3})}{\text{var}(x_{1.3})}$ i.e. $r_{21.3} \sigma_{2.3} / \sigma_{1.3}$ etc. Generalizing this trivariate case to that of a p -variates, we

can show that $b_{12.q} = \frac{\Sigma x_{1.q} x_{2.q}}{\Sigma x_{2.q}^2}, \text{ or } \frac{\text{cov}(x_{1.q}, x_{2.q})}{\text{var}(x_{2.q})}$ i.e. $r_{12.q} \sigma_{1.q} / \sigma_{2.q}$,

where q stands for the group of suffixes 3, 4, ..., p . Thus we observe that a regression coefficient of order $(p-2)$ is expressible in terms of σ , and r , each of order $(p-2)$.

We, therefore, arrive at the conclusion that a partial regression coefficient can be expressed in terms of s.d.s. and correlation coefficients of zero orders as well as of higher orders.

Exp. (10). [a] Show that the coefficient of correlation lies between -1 and $+1$.

[b] For given r_{12} , r_{13} find the range of r_{23} and also comment on $r_{12.3} = 0, 1$.

[c]. (i) If $r_{12} = k$, $r_{23} = -k$, prove that $-1 \leq r_{13} \leq 1 - 2k^2$.

(ii) If $r_{23} = 0$, prove that $R^2_{1(23)} = r^2_{12} + r^2_{13}$; and $\sigma^2_{1.23} = \sigma^2_1(1 - r^2_{12} - r^2_{13})$.

(iii) If $r_{23} = 1$, prove that $r^2_{12} = r^2_{13}$; and $\sigma^2_{1.23} = \sigma^2_1(1 - r^2_{12})$.

Sol. [a] Let r be the correlation coefficient between x and y for a random sample of n pairs drawn from a bivariate normal population. If \bar{x} , \bar{y} denote the sample means and σ_x , σ_y denote the sample s.d.s. of the variables x and y , then by the property of sum of squares, we have

$$S_1 = \frac{1}{2n} \Sigma \left[\frac{x - \bar{x}}{\sigma_x} - \frac{y - \bar{y}}{\sigma_y} \right]^2 \geq 0$$

$$\text{or } \frac{1}{2} \left[\frac{\Sigma (x - \bar{x})^2}{n\sigma_x^2} + \frac{\Sigma (y - \bar{y})^2}{n\sigma_y^2} - 2 \frac{\Sigma (x - \bar{x})(y - \bar{y})}{n\sigma_x\sigma_y} \right] \geq 0$$

$$\text{or } \frac{1}{2} \left[\frac{n\sigma_x^2}{n\sigma_x^2} + \frac{n\sigma_y^2}{n\sigma_y^2} - 2 \frac{\text{cov}(x, y)}{\sigma_x\sigma_y} \right] \geq 0$$

$$\text{or } \frac{1}{2} [1 + 1 - 2r] \geq 0$$

$$\text{or } 1 - r \geq 0, \therefore r \leq +1 \dots \dots \dots (1)$$

Similarly, by considering the quantity

$$S_2 = \frac{1}{2n} \sum \left[\frac{x - \bar{x}}{\sigma_x} + \frac{y - \bar{y}}{\sigma_y} \right]^2 \geq 0, \text{ we have } 1 + r \geq 0$$

$$\therefore r \geq -1 \dots \dots \dots (2)$$

Combining the two results obtained in (1), (2) we see that $-1 \leq r \leq +1$. Thus the coefficient of correlation between any two variables always lies between -1 and $+1$. **Hence proved.**

[b] We know that

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}} \leq 1$$

$$\text{or } r_{12.3}^2 = \frac{(r_{12} - r_{13}r_{23})^2}{(1-r_{13}^2)(1-r_{23}^2)} \leq 1$$

$$\text{i.e. } (r_{12} - r_{13}r_{23})^2 \leq (1-r_{13}^2)(1-r_{23}^2)$$

$$\text{or } r_{23}^2 - 2r_{12}r_{13}r_{23} + (r_{13}^2 + r_{12}^2 - 1) \leq 0 \dots \dots \dots (1)$$

If r_{12} , r_{13} are known, then the equ. (1) is analogous to the quadratic equ. $ax^2 + bx + c \leq 0$ giving $x = \frac{-b \pm \sqrt{(b^2 - 4ac)}}{2a}$, where

a , b and c are known constants.

Thus from (1), we have

$$r_{23} = \frac{2r_{12}r_{13} \pm \sqrt{[4r_{12}^2r_{13}^2 - 4(r_{12}^2 + r_{13}^2 - 1)]}}{2}$$

$$\text{or } r_{23} = r_{12}r_{13} \pm \sqrt{(r_{12}^2r_{13}^2 - r_{12}^2 - r_{13}^2 + 1)} \dots \dots \dots (2)$$

Thus the limits of r_{23} will be $r_{12}r_{13} \pm \sqrt{(r_{12}^2r_{13}^2 - r_{12}^2 - r_{13}^2 + 1)}$ giving its range from $[r_{12}r_{13} - \sqrt{(r_{12}^2r_{13}^2 - r_{12}^2 - r_{13}^2 + 1)}]$ to $[r_{12}r_{13} + \sqrt{(r_{12}^2r_{13}^2 - r_{12}^2 - r_{13}^2 + 1)}]$.

Further we see that if $r_{12.3} = 0$, then $\frac{r_{12} - r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}} = 0$, which is only possible when $r_{12} = 0$, and r_{13} (or $r_{23}) = 0$. It means that x_1 (or x_2) is uncorrelated with x_2 (or x_1) and x_3 . This gives us the *necessary and sufficient*(n&s) condition for the uncorrelated variables of the three regression planes for a trivariate normal distribution.

Also, if $r_{12.3} = 1$, then $\frac{r_{12} - r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}} = 1$ i.e. $r_{12}^2 + r_{23}^2 + r_{31}^2 - 2r_{12}r_{23}r_{31} = 1$, which is only possible when the three planes coincide. This gives us the *n & s* condition for the coincidence of the three regression planes for a trivariate normal distribution.

[c].(i) If r_{12} , r_{23} are known, then the limits of r_{13} will be :

$$r_{12}r_{23} \pm \sqrt{(r_{12}^2r_{23}^2 - r_{12}^2 - r_{23}^2 + 1)}, \text{ as obvious from (2) of [b].}$$

$$\text{or } k(-k) \pm \sqrt{(k^4 - k^2 - k^2 + 1)} \quad (\text{since } r_{12} = k, r_{23} = -k)$$

$$\text{or } -k^2 \pm \sqrt{(1-k^2)^2}; \text{ or } -k^2 \pm (1-k^2)$$

$$\text{i.e. } -k^2 - (1-k^2) \text{ and } -k^2 + (1-k^2); \text{ or } -1 \text{ and } 1-2k^2,$$

$$\therefore -1 \leq r_{13} \leq 1-2k^2,$$

Hence proved.

(ii) Since we know that $R_{1(23)}$ and r are related by the equation

$$R_{1(23)}^2 = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{23}r_{31}}{1 - r_{23}^2},$$

so $R_{1(23)}^2 = r_{12}^2 + r_{13}^2$, when $r_{23} = 0$.

Also, $1 - R_{1(23)}^2 = 1 - (r_{12}^2 + r_{13}^2)$; or $\sigma_{1.23}^2/\sigma_1^2 = (1 - r_{12}^2 - r_{13}^2)$

$\therefore \sigma_{1.23}^2 = \sigma_1^2(1 - r_{12}^2 - r_{13}^2)$. **Hence proved.**

(iii) As we know that

$$R_{1(23)}^2(1 - r_{23}^2) = r_{12}^2 + r_{13}^2 - 2r_{12}r_{23}r_{31},$$

so $0 = r_{12}^2 + r_{13}^2 - 2r_{12}r_{31}$, when $r_{23} = 1$.

or $0 = (r_{12} - r_{13})^2$

$\therefore r_{12} = r_{13}$, and consequently $r_{12}^2 = r_{13}^2$.

Aliter : Since we know that

$$r_{23} = r_{12}r_{13} \pm \sqrt{(r_{12}^2r_{13}^2 - r_{12}^2 - r_{13}^2 + 1)},$$

so $(r_{23} - r_{12}r_{13})^2 = (r_{12}^2r_{13}^2 - r_{12}^2 - r_{13}^2 + 1)$

or $(1 - r_{12}r_{13})^2 = (r_{12}^2r_{13}^2 - r_{12}^2 - r_{13}^2 + 1)$ when $r_{23} = 1$.

or $(r_{12} - r_{13})^2 = 0$, i.e. $r_{12} = r_{13}$,

Hence proved.

Further, as we know that

$$\begin{aligned} \sigma_{1.23}^2 &= \sigma_1^2(1 - r_{12}^2)(1 - r_{13.2}^2) \\ &= \sigma_1^2(1 - r_{12}^2) \left[1 - \frac{(r_{13} - r_{12}r_{23})^2}{(1 - r_{12}^2)(1 - r_{23}^2)} \right] \end{aligned}$$

$\therefore \sigma_{1.23}^2 = \sigma_1^2(1 - r_{12}^2)$, when $r_{23} = 1$. **Hence proved.**

Exp. (11). [a] Show that $R_{1(23)}^2 = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{23}r_{31}}{1 - r_{23}^2} \geq r^2$.

Hence or otherwise establish the result $\sigma_{1.23 \dots p}^2/\sigma_1^2 = (1 - r_{12}^2)(1 - r_{13.2}^2)(1 - r_{14.23}^2) \dots (1 - r_{1p.23 \dots p-1}^2) = 1 - R_{1(23 \dots p)}^2$.

[b] Prove that $R_{1(23)}^2 = b_{12.3} \frac{\sigma_2}{\sigma_1} + b_{13.2} \frac{\sigma_3}{\sigma_1}$.

[c] (i) Show that $b_{13.2} = (b_{12} - b_{13}b_{32})/(1 - b_{23}b_{32})$.

(ii) ,, ,, $r_{12.3} \frac{\sigma_{1.3}}{\sigma_{2.3}} = \frac{\sigma_1}{\sigma_2} \left(\frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2} \right)$.

(iii) ,, ,, $b_{12.3}b_{23.1}b_{31.2} = r_{12.3}r_{23.1}r_{31.2}$.

Sol. [a] We know that

$$\begin{aligned} 1 - R_{1(23)}^2 &= \frac{\Delta}{\Delta_{11}} = \frac{\Delta_{11} + r_{12}\Delta_{12} + r_{13}\Delta_{13}}{\Delta_{11}} \\ &= \frac{(1 - r_{23}^2) + r_{12}(r_{31}r_{32} - r_{12}) + r_{13}(r_{12}r_{23} - r_{31})}{(1 - r_{23}^2)} \\ &= \frac{1 - r_{23}^2 - r_{12}^2 - r_{13}^2 + 2r_{12}r_{23}r_{31}}{1 - r_{23}^2} \dots \dots \dots (1) \\ &= 1 - \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{23}r_{31}}{1 - r_{23}^2} \end{aligned}$$

$$\therefore R_{1(23)}^2 = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{23}r_{31}}{1 - r_{23}^2} \dots \dots \dots (2)$$

Further by adding and subtracting the quantity $r_{12}^2 r_{23}^2$ in the numerator of (1), we have

$$\begin{aligned} 1 - R_{1(23)}^2 &= \frac{1 - r_{23}^2 - r_{12}^2 + r_{12}^2 r_{23}^2 - r_{12}^2 r_{23}^2 - r_{12}^2 + 2r_{12} r_{23} r_{31}}{1 - r_{23}^2} \\ &= \frac{(1 - r_{12}^2)(1 - r_{23}^2) - (r_{12}^2 r_{23}^2 + r_{12}^2 - 2r_{12} r_{23} r_{31})}{1 - r_{23}^2} \\ &= \frac{(1 - r_{12}^2)}{(1 - r_{12}^2)} \times \frac{(1 - r_{12}^2)(1 - r_{23}^2) - (r_{12}^2 - r_{12} r_{23} r_{31})^2}{(1 - r_{23}^2)} \\ &= (1 - r_{12}^2) \left[\frac{(1 - r_{12}^2)(1 - r_{23}^2)}{(1 - r_{12}^2)(1 - r_{23}^2)} - \frac{(r_{12}^2 - r_{12} r_{23} r_{31})^2}{(1 - r_{12}^2)(1 - r_{23}^2)} \right] \end{aligned}$$

$$\therefore 1 - R_{1(23)}^2 = (1 - r_{12}^2)(1 - r_{12}^2) \dots \dots \dots (3)$$

Thus from (3), it is obvious that $1 - R_{1(23)}^2 \leq 1 - r_{12}^2$; i.e. $R_{1(23)}^2 \geq r_{12}^2$;
or $R_{1(23)}^2 \geq r_{12}^2 \dots \dots \dots (4)$

Combining the two results obtained in (2), (4) we see that

$$R_{1(23)}^2 = \frac{r_{12}^2 + r_{12}^2 - 2r_{12} r_{23} r_{31}}{1 - r_{23}^2} \geq r_{12}^2. \quad \text{Hence proved.}$$

Again considering the relation (3), we have

$$\sigma_{1,23}^2 / \sigma_1^2 = 1 - R_{1(23)}^2 = (1 - r_{12}^2)(1 - r_{12}^2).$$

Generalizing this trivariate case to that of a p-variables, we have

$$\sigma_{1,23\dots p}^2 / \sigma_1^2 = (1 - r_{12}^2)(1 - r_{13}^2)(1 - r_{14}^2) \dots \dots (1 - r_{1p}^2) \dots \dots = 1 - R_{1(23\dots p)}^2.$$

Hence the result.

[b] Since we know that

$$\begin{aligned} R_{1(23)}^2 &= 1 - \frac{\Delta}{\Delta_{11}} = 1 - \frac{\Delta_{11} + r_{12} \Delta_{12} + r_{13} \Delta_{13}}{\Delta_{11}} \\ &= 1 - \left[1 + r_{12} \frac{\Delta_{12}}{\Delta_{11}} + r_{13} \frac{\Delta_{13}}{\Delta_{11}} \right] \\ &= r_{12} \left(-\frac{\Delta_{12}}{\Delta_{11}} \right) + r_{13} \left(-\frac{\Delta_{13}}{\Delta_{11}} \right) \\ &= r_{12} \left(\frac{\sigma_2}{\sigma_1} b_{2,3} \right) + r_{13} \left(\frac{\sigma_3}{\sigma_1} b_{3,2} \right), \text{ since } b_{12,3} = -\frac{\sigma_1}{\sigma_2} \cdot \frac{\Delta_{12}}{\Delta_{11}} \text{ etc.} \end{aligned}$$

$$\text{So } R_{1(23)}^2 = b_{12,3} r_{12} \frac{\sigma_2}{\sigma_1} + b_{13,2} r_{13} \frac{\sigma_3}{\sigma_1}. \quad \text{Hence proved.}$$

[c]. (i) We know that

$$\begin{aligned} b_{12,3} &= -\frac{\sigma_1}{\sigma_2} \cdot \frac{\Delta_{12}}{\Delta_{11}} \\ &= -\frac{\sigma_1}{\sigma_2} \cdot \frac{r_{12} r_{23} - r_{13}}{1 - r_{23}^2} = -\frac{\sigma_1}{\sigma_2} \frac{r_{12} r_{23} + \frac{\sigma_1}{\sigma_2} r_{12}}{1 - b_{23} b_{32}} \\ &= -\frac{\sigma_1}{\sigma_2} r_{12} \times \frac{\sigma_2}{\sigma_2} r_{23} + b_{12} = -\frac{b_{12} b_{23} + b_{12}}{1 - b_{23} b_{32}} = -\frac{b_{12} b_{23} + b_{12}}{1 - b_{23} b_{32}} \end{aligned}$$

$$\therefore b_{12,3} = (b_{12} - b_{12} b_{32}) / (1 - b_{23} b_{32}). \quad \text{Hence proved.}$$

(ii) We have

$$\begin{aligned}
 r_{12 \cdot 3} \frac{\sigma_{1 \cdot 3}}{\sigma_{2 \cdot 3}} &= b_{12 \cdot 3} \\
 &= \frac{-\sigma_1}{\sigma_2} \cdot \frac{\Delta_{12}}{\Delta_{11}} = \frac{-\sigma_1}{\sigma_2} \cdot \frac{(r_{13}r_{23} - r_{12})}{(1 - r_{12}^2)} \\
 \therefore r_{12 \cdot 3} \frac{\sigma_{1 \cdot 3}}{\sigma_{2 \cdot 3}} &= \frac{\sigma_1}{\sigma_2} \left(\frac{r_{12} - r_{13}r_{23}}{1 - r_{12}^2} \right). \quad \text{Hence proved.}
 \end{aligned}$$

(iii) We know that

$$\begin{aligned}
 b_{12 \cdot 3} b_{23 \cdot 1} b_{31 \cdot 2} &= \frac{\text{COV}(x_{1 \cdot 3}, x_{2 \cdot 3})}{\text{var}(x_{2 \cdot 3})} \times \frac{\text{COV}(x_{2 \cdot 1}, x_{3 \cdot 1})}{\text{var}(x_{3 \cdot 1})} \times \frac{\text{COV}(x_{3 \cdot 2}, x_{1 \cdot 2})}{\text{var}(x_{1 \cdot 2})} \\
 &= \frac{r_{12 \cdot 3} \sigma_{1 \cdot 3} \sigma_{2 \cdot 3}}{\sigma_{2 \cdot 3}^2} \times \frac{r_{23 \cdot 1} \sigma_{2 \cdot 1} \sigma_{3 \cdot 1}}{\sigma_{3 \cdot 1}^2} \times \frac{r_{31 \cdot 2} \sigma_{3 \cdot 2} \sigma_{1 \cdot 2}}{\sigma_{1 \cdot 2}^2} \\
 &= \frac{r_{12 \cdot 3} \sigma_{1 \cdot 3}}{\sigma_{2 \cdot 3}} \times \frac{r_{23 \cdot 1} \sigma_{2 \cdot 1}}{\sigma_{3 \cdot 1}} \times \frac{r_{31 \cdot 2} \sigma_{3 \cdot 2}}{\sigma_{1 \cdot 2}} \\
 &= r_{12 \cdot 3} \frac{\sigma_1 \sqrt{(1 - r_{13}^2)}}{\sigma_2 \sqrt{(1 - r_{12}^2)}} \times r_{23 \cdot 1} \frac{\sigma_2 \sqrt{(1 - r_{21}^2)}}{\sigma_3 \sqrt{(1 - r_{23}^2)}} \times r_{31 \cdot 2} \frac{\sigma_3 \sqrt{(1 - r_{32}^2)}}{\sigma_1 \sqrt{(1 - r_{31}^2)}}
 \end{aligned}$$

$$\therefore b_{12 \cdot 3} b_{23 \cdot 1} b_{31 \cdot 2} = r_{12 \cdot 3} r_{23 \cdot 1} r_{31 \cdot 2}.$$

Hence proved.

Exp. (12). [a] Show that the values : $r_{12}=0.6$, $r_{23}=0.8$, and $r_{31}=-0.5$ are inconsistent.

[b] In a sample of 28 from a trivariate normal population, it has been found that $r_{12}=.8$, $r_{13}=-.4$ and $r_{23}=-.56$. Compute $r_{12 \cdot 3}$, $R_{1(23)}$, and test their significance.

[c] From independent samples of 31 and 22 sets of values, partial correlations of order three are found to be .4 and .6 respectively. Examine

(i) whether the first sample could have come from a population with the corresponding correlation coefficient of 0.6;

(ii) whether the two samples could have come from the same normal population.

Sol. [a] If a partial correlation coefficient computed from the given values $r_{12}=0.6$, $r_{23}=0.8$, and $r_{31}=-0.5$, lies between -1 and $+1$, then only the values are said to be consistent, otherwise inconsistent. Let us compute, for the purpose, the partial correlation coefficient $r_{12 \cdot 3}$ as follows--

$$\begin{aligned}
 r_{12 \cdot 3} &= \frac{r_{12} - r_{13} r_{23}}{\sqrt{[(1 - r_{13}^2)(1 - r_{23}^2)]}} \\
 &= \frac{0.6 - (-0.5)(0.8)}{\sqrt{[(1 - .25)(1 - .64)]}} = \frac{0.60 + 0.40}{\sqrt{[(.75)(.36)]}} = \frac{1.0}{5.196} > 1,
 \end{aligned}$$

hence inadmissible.

Conclusion : The given values are inconsistent.

$$[b] \text{ We have } r_{12 \cdot 3} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{[(1 - r_{13}^2)(1 - r_{23}^2)]}} = \frac{0.8 - (-.4)(-.56)}{\sqrt{[(1 - .16)(1 - .31)]}} = 0.76$$

$$\text{and } R^2_{1(23)} = \frac{r^2_{12} + r^2_{13} - 2r_{12}r_{23}r_{31}}{1 - r^2_{23}} = \frac{.64 + .16 - 2(.8)(-.4)(-.56)}{1 - .31} = 0.64, \therefore R_{1(23)} = .8.$$

For testing the significance of $r_{12.3}$ ($=0.76$), we take the hyp. (H₀) : $\rho=0$.

Here we compute the statistic

$$t = \frac{r}{\sqrt{[(1-r^2)/(n-k-2)]}} = \frac{0.76}{\sqrt{[(1-(.76)^2)/(28-1-2)]}} = \frac{0.76 \times 5}{.4224} = 8.99$$

Now we have $|t| = 8.99$ and $t_{.05}(25) = 2.06$.

So $|t| > t_{.05}$ leading to the rejection of hyp. at 5% level.

Similarly for testing the significance of $R_{1(23)} = .80$, we take the hyp. (H₀) : $\bar{R}=0$.

Here we compute the statistic

$$F = \frac{R^2}{1-R^2} \cdot \frac{n-p-1}{p} = \frac{0.64}{1-0.64} \cdot \frac{28-2-1}{2} = \frac{0.64 \times 25}{0.36 \times 2} = 22.22.$$

Now we have $F = 22.22$ and $F_{.05}(2, 25) = 3.38$.

So $F > F_{.05}$ leading to the rejection of hyp. at 5% level.

Conclusion : The computed values of both $r_{12.3}$ and $R_{1(23)}$ are significant at 5% level of significance.

[c]. (i) Here we are given : $r=0.4$, $k=3$, $n=31$; and want to test the hyp. (H₀) : $\rho=0.6$.

Thus we compute the statistic

$$Z = \frac{z - \xi}{1/\sqrt{(n-k-3)}}, \text{ where } z = 1.1513 \log_{10} \left(\frac{1+r}{1-r} \right) = .42, \text{ and}$$

$$\xi = 1.1513 \log_{10} \left(\frac{1+\rho}{1-\rho} \right) = .69.$$

$$= \frac{.42 - .69}{1/\sqrt{(31-3-3)}} = \frac{-.27}{1} \times 5 = -1.35.$$

So $|Z| < 1.96$ giving no evidence against the hyp. at 5% level.

Conclusion : The first sample with observed partial correlation coefficient of 0.4 may be supposed as drawn from a population with corresponding correlation coefficient of 0.6.

(ii) Here we have been given : $r_1=0.4$, $n_1=31$, $r_2=0.6$, $n_2=22$, $k=3$, and want to test the hyp. (H₀) : $\rho_1=\rho_2$.

Thus we compute the statistic

$$Z = \frac{z_1 - z_2}{\sqrt{\left(\frac{1}{n_1 - k - 1} + \frac{1}{n_2 - k - 3}\right)}}, \text{ where } z_1 = 1.1513 \log_{10} \left(\frac{1+r_1}{1-r_1} \right) = .42, \text{ and } z_2 = .69$$

$$= \frac{.42 - .69}{\sqrt{\left(\frac{1}{31-3-3} + \frac{1}{22-3-3}\right)}} = \frac{-.27}{\sqrt{\left(\frac{1}{25} + \frac{1}{16}\right)}} = \frac{-.54}{6.4} = -.84$$

So $|Z| < 1.96$ giving no evidence against the hyp. at 5% level.

Conclusion : The two independent samples with observed partial correlation coefficients of 0.4 and 0.6 may be regarded as drawn from the same normal population.

Exp (13) Given that $\bar{x}_1=8$, $\bar{x}_2=5$, $\bar{x}_3=4$, $r_{12}=0.86$, $r_{23}=0.72$, $r_{31}=0.65$, $\sigma_1=0.1204$, $\sigma_2=0.1306$, and $\sigma_3=0.154$. Set out the multiple regression equation of x_1 on x_2 , x_3 and find x_1 for $x_2=8$, $x_3=6$. Also find the s.e. of estimate of x_1 .

Sol. The multiple regression equation of x_1 on x_2 , x_3 is

$$\frac{x_1 - \bar{x}_1}{\sigma_1} \Delta_{11} + \frac{x_2 - \bar{x}_2}{\sigma_2} \Delta_{12} + \frac{x_3 - \bar{x}_3}{\sigma_3} \Delta_{13} = 0$$

$$\text{or } \frac{x_1 - 8}{0.1204} (0.4816) + \frac{x_2 - 5}{0.1306} (-0.3920) + \frac{x_3 - 4}{0.1540} (-0.0308) = 0$$

$$\text{or } (x_1 - 8)(4) + (x_2 - 5)(-3) + (x_3 - 4)(-0.2) = 0$$

$$\begin{cases} \text{since } \Delta_{11} = (1 - r_{23}^2) = .4816 \\ \Delta_{12} = (r_{31}r_{32} - r_{12}) = -.392 \\ \Delta_{13} = (r_{12}r_{23} - r_{31}) = -.0308 \end{cases}$$

$$\text{or } 4x_1 - 32 - 3x_2 + 15 - 0.2x_3 + 0.8 = 0$$

$$\text{i.e. } x_1 = 0.5x_2 + 0.5x_3 + 4.05.$$

Hence the required equation.

$$\text{Also, } x_1 = 0.75(8) + 0.05(6) + 4.05, \text{ when } x_2=8, x_3=6.$$

$$\therefore x_1 = 10.35.$$

Hence the required estimate.

Now we have

$$\begin{aligned} \text{S.E.}(x_1) &= \sigma_{1.23} = \sigma_1 \left(\frac{\Delta}{\Delta_{11}} \right)^{1/2} = \sigma_1 \left[\frac{\Delta_{11} + r_{12}\Delta_{12} + r_{13}\Delta_{13}}{\Delta_{11}} \right]^{1/2} \\ &= 0.1204 \left[\frac{0.4816 + 0.86(-0.3920) + 0.65(-0.0308)}{0.4816} \right]^{1/2} \\ &= 0.1204 \left[\frac{0.1245}{0.4816} \right]^{1/2} = 0.1204 (0.25)^{1/2} = 0.1204 \times .5 \end{aligned}$$

$$\therefore \text{S.E.}(x_1) = 0.0602.$$

Hence the result.

EXERCISE VII

1. A research worker observed the following fresh and dry weights in ounces for a sample of his experimental material —

Fresh wts... 8 6 10 5 12 2 20 15 14 18

Dry wts ... 3 2 2 2 4 1 5 4 3 4.

[a] Calculate the correlation between the two characters.

[b] Represent the data in a scatter diagram. (M.Sc. Ag AU, 1963)

2. Find the two regression lines and also plot them on a graph to give an idea of correlation between x and y for the following —

x .. 25 27 26 29 34 35

y ... 2 5 7 9 19 17.

3. The following data are for the amount of water applied in inches and yield of alfalfa in tons/acre—

Water (x) ..12 18 24 30 36 42 48

Yield (y) ..5.27 5.68 6.25 7.21 8.20 8.67 8.42. Find the regression of yield on water. Assuming that the relation between the two is linear, find the expected yield when the amount of water applied is 20".

4. The following table gives the temperature of unhusked rice and the % breakage of rice grain in milling —

Temp. (degree)... 33 34 29 35 38 28 29 36 34 30

% breakage ... 37 24 26 27 30 24 25 28 30 24.

Calculate the coefficient of correlation between the temperature and % breakage, and also test its significance. (M.Sc.Ag. AU, 1961)

5. Define the coefficient of correlation and find its value between x and y for the data given below—

x ... 8 10 15 17 20 22 24 25

y ... 25 30 32 35 37 40 42 45.

6. What do you understand by perfect positive and perfect negative correlations? Is a significant test needed for this value? The correlation coefficient between *days to head* and *days to mature* for varieties was found to be 0.4, is this significant?

7. The following table enumerates the marks obtained by a class of students in Statistics in 1st and 2nd papers—

Marks in 1st (x)...80 45 55 56 58 60 65 68 70 75 85

„ „ 2nd (y)...82 56 50 48 60 62 64 65 70 74 90.

Calculate the two regression coefficients and hence or otherwise compute the correlation coefficient between x and y .

8. A sample of paired variates is given below—

x : 9 8 7 7 6 5 3 3 1 1

y : 9 9 8 6 6 5 4 3 1 1.

[a] Calculate r between x and y , and interpret the same.

[b] Illustrate the scatter of the points in a diagram.

(M. Sc. Ag AU, 1965)

9. From the regression lines : $10y - 16x - 21 = 0$

and $2y - 5x + 16 = 0$.

[a] Find the mean of x and y .

[b] Compute the coefficient of correlation between x and y .

[c] Plot the lines on a graph and give an idea of correlation.

Also, obtain the means \bar{x} and \bar{y} from the graph and compare them with those of obtained in [a]

10 For six pairs of observations of x and y , the following deviated values were calculated—

$\Sigma \xi = -1, \Sigma \xi^2 = 229, \Sigma \eta = -4, \Sigma \eta^2 = 92$, and $\Sigma \xi \eta = 139$, where ξ, η stand for the deviations of x, y respectively from the corresponding assumed means.

Compute r and b_{yx} ; and also test their significance.

11. [a] In a sample of 20 pairs of values, *total solids* and *fat contents* gave the correlation coefficient of 0.6. What would you conclude from this value? (M. Sc. Ag. AU, 1959)

[b] The thickness of 20 *annual rings* of a tree and the corresponding *annual rainfall* were found to be correlated with a coefficient of +0.47. Is this correlation significant?

[c] For two characters x and y , it is found that : $r = +0.8$, $\bar{x} = 25$, $\bar{y} = 22$, $\sigma_x = 4$ and $\sigma_y = 5$.

Calculate (i) the expected value of x for $y = 12$, and

(ii) the expected value of y for $x = 33$ (M. Sc. Ag. AU, 1962)

12. [a] The ranks of 15 participants in a beauty contest graded by two judges were as follows—

(6,8);(7,3);(8,1);(9,11);(10,15);(11,9);(12,5);(13,14);(14,12);(15,13)
(1,10);(2,7);(3,2);(4,6);(5,4). Show that the two judges do not differ in their opinions.

[b] The ranks of the same 16 students in two subjects A and B were as follows—

(1,1),(2,10),(3,3),(4,4),(5,5),(6,7),(7,2),(8,6),(9,8),(10,11),(11,15),
(12,9),(13,14),(14,12),(15,16),(16,13). Calculate the rank correlation for proficiencies of this group in subject A and B.

13. [a] The following marks, out of 100, have been obtained by a class of ten students in Statistics—

Student ... I II III IV V VI VII VIII IX X

Paper I ... 78 60 38 75 45 25 85 89 52 32

Paper II... 45 65 34 35 82 76 72 30 56 42. Compute the coefficient of rank correlation and comment on the proficiencies of the class in the two papers.

[b] Ten competitors in a beauty contest are ranked by three judges in the following orders—

1st judge ... 1 6 5 10 3 2 4 9 7 8

2nd „ ... 3 5 8 4 7 10 2 1 6 9

3rd „ ... 6 4 9 8 1 2 3 10 5 7 Use the rank correlation coefficient to discuss which pair of judges have the nearest approach to common tastes in beauty.

14. In a sample of 20 sets of values of three mutually dependent variables x_1 , x_2 and x_3 , the following data are available—

$$\left. \begin{array}{l} \bar{x}_1=20 \\ \bar{x}_2=15 \\ \bar{x}_3=18 \end{array} \right\} \left. \begin{array}{l} \sigma_1=2.0 \\ \sigma_2=1.5 \\ \sigma_3=3.0 \end{array} \right\} , \text{ and } \left. \begin{array}{l} r_{12}=0.4 \\ r_{23}=-0.2 \\ r_{31}=0.3 \end{array} \right\} . \text{ Calculate (i) any three partial regression coefficients, (ii) the}$$

multiple correlation coefficient of x_1 on x_2 , x_3 ; and also test its significance; (iii) the estimate of x_3 for $x_2=12$, $x_1=15$; and the s.e. of the estimate.

15. In an experimental study on Cinchona plants, the following quantities were obtained from a sample of 32 plants—

$$\left. \begin{array}{l} \bar{x}_1=21.68 \\ \bar{x}_2=166.4 \\ \bar{x}_3=3.14 \end{array} \right\} \left. \begin{array}{l} \sigma_1=14.25 \\ \sigma_2=56.67 \\ \sigma_3=1.03 \end{array} \right\} \left. \begin{array}{l} r_{12}=.367 \\ r_{23}=.321 \\ r_{31}=.684 \end{array} \right\} , \text{ where } x_1 \text{ represents the yield of bark (oz.), } \bar{x}_2 \text{ the height of plants (inches) and } x_3 \text{ the girth}$$

of plants 6" above the ground (inches). Calculate (i) $r_{12.3}$, $b_{12.3}$ and $R_{2(13)}$. (ii) Set out the multiple regression equation of x_1 on x_2 , x_3 ; and find the most likely value of x_1 when $x_2=150$ and $x_3=4$. Also find the s.e. of the estimate.

ANSWERS

(1) $r = +0.91$ (2) $y = 1.55x - 35.63$; $x = 0.60y + 23.43$; limited (+)ve correlation (3) $y = 0.1035x + 5.995$, $y_e = 8.135$ tons/acre (+) $r = +0.49$, $t = 1.48$ (5) $r = +0.98$ (6) $r = \pm 1$, no, $t = .23$ (7) $b_{yx} = 0.99$, $b_{xy} = 0.85$, $r = +0.92$ (8) $r = +0.98$, high (+)ve; (9). [a] $\bar{x} = 11.22$, $\bar{y} = 20.05$ [b] $r = +0.8$ (10) $r = +0.96$, $t = 7.38$ $b_{yx} = 0.60$, $t = 7.5$ (11). [a] $t = 3.18$ [b] $t = 2.26$ [c]. (i) $x_e = 18.60$ (ii) $y_e = 30$ (12). [a] $r = +0.51$, the positive value of rank correlation coefficient shows that the two judges do not differ in their opinions. [b] $r = +0.8$ (13). [a] $r = -0.3$, the standards of ability of the students are different in the two papers [b] $r_{12} = +0.7$, $r_{13} = 0 = r_{23}$ thus the last two pairs of judges have the least disparity in their approach to beauty. (14). (i) $b_{12.3} = 0.674$, $b_{23.1} = 0.198$, $b_{13.2} = -0.234$, (ii) $R_{1(23)} = 0.512$, $F = 3.14$, (iii) $x_{3e} = 18.14$, $\sigma_{3.12} = 2.66$ (15) (i) $r_{12.3} = 0.2134$, $b_{12.3} = 0.041$, $R_{2(13)} = 0.38$ (ii) $x_{1e} = 0.0413x_2 + 8.733x_3 - 12.6139$, $x_{1e} = 28.5131$, $\sigma_{1.23} = 10.15$.

Chapter VIII

Sampling Techniques

8.1 Introduction : Before designing any plan, we essentially require some quantitative data which can be obtained through an enquiry or survey. This survey may be of two types—(a) *census survey*, and (b) *sample survey*. A census survey is one in which all the units connected with the problem of investigation are taken into account, while in a sample survey only some selected units are considered.

8.1 (a) Census survey : Here the conclusions regarding the population are based on the complete enumeration of the whole experimental material and hence the technique requires a lot of time expenditure and other necessary resources. Thus the use of the technique is limited and it can be applied only to the problems related to some special fields of enquiry such as the census count, production, imports and exports etc. of certain principal commodities.

Below we give some specific situations in which the technique can successfully be employed—

- (i) When the population or field of investigation is limited.
- (ii) When the results are needed with maximum possible accuracy and reliability.
- (iii) When the nature of the units under investigation is less dynamic,
- (iv) When the selection of a sample is difficult.
- (v) When the sufficient money, time and other resources are available.

(a-1) Limitations of a census survey :

(i) It cannot be applied to an infinite population, or a population which is not completely known.

(ii) It does not give us a method of knowing the precision of the results drawn.

(iii) It cannot be used in the situations wherein the units to be collected have the considerable variation among themselves for the character under study.

(iv) It cannot be applied to the problems in which the sampling is easy, or the study is destructive.

(v) It requires a lot of money, time, manpower and other resources.

8.1 (b) Sample survey : Here the conclusions regarding the parent population are based on the results obtained only from a few selected units of the population and hence the technique saves a considerable time, expenditure and other necessary resources. Thus the technique has a greater scope and a vast-field of application to different types of enquiries such as the estimation of yield, area and income etc. of certain crops.

Below we give some specific situations in which the 'technique can successfully be employed—

(i) When the population or field of investigation is infinite, or the population is not completely known.

(ii) When a character is to be estimated with some specified precision.

(iii) When the units to be collected have a considerable variation among themselves for the character under study.

(iv) When the complete enumeration is not possible, or the study is destructive.

(v) When the time, money and other resources are limited.

(b-1) Advantages of a sample survey : The main advantages of a sample survey in comparison to a census survey are as follows—

(i) **Greater adaptability :** In most of the cases where the population is infinite, or the whole of the population material is not known, or a destructive test is applied to measure the character under study, then it is impossible to collect the information regarding the whole population. Thus in all such cases, a sample survey is the only scientific method to be adopted for the purpose. Also there is no alternative except sampling to know the precision of the results obtained.

(ii) **Greater scope :** In some enquiries where an intensive study is to be made and the highly trained personnel or the specialized equipments are scarcely available, a sample survey gives the more scope of enquiries than with a census survey.

(iii) **Greater speed :** The data can be collected and interpreted more quickly with a sample than with a census.

(iv) **Greater economy :** The study of the population through a sample saves a considerable amount of cost, as only a small fraction of the aggregate is considered here.

(v) **Greater accuracy :** By employing the trained personnel that are always available only in a limited number, the results of a sample survey may be more accurate than a census survey.

(b-2) Limitations of a sample survey :

(i) In spite of the fact that a proper method of selection is

employed, a sample does not truly represent its parent population and consequently the results are less reliable.

(ii) The problems of choosing a proper sampling unit, size and technique are too difficult and technical.

(iii) There is always a possibility of the inclusion of a human bias.

(iv) It requires a good background of mathematics needed for computing the estimates and their s.e.s.

(b-3) Main steps of a sample survey : The requisite steps in designing a sample survey are as follows---

(i) A clear statement of the objective of survey.

(ii) The definition and size of the parent population to be sampled.

(iii) The choice of a measuring device.

(iv) „ „ „ „ sampling unit.

(v) „ „ „ „ „ technique.

(vi) „ „ „ „ sample size.

(vii) The plan and organization of the field work.

(viii) The description of the expenditure, time and other resources to be utilized in the survey.

(ix) The summary and analysis of the collected data.

(x) The information availed for future survey.

8.2 Technical terms of a sample survey : Before any further discussion of the sampling theory, we think it more illustrative and advantageous to a beginner to define precisely some main technical terms that are commonly used in the sampling theory.

(i) **Population :** *In statistics, the whole field of investigation is called a population or universe.* It may also be termed as the collection of individuals or of their attributes numerically specified. Further, a population composed of real individuals is called an *existent population*, while an aggregate of all feasible ways in which an event can happen is called a *hypothetical population*. The total no. of plants in a field is an example of a real or existent population, while the total no. of heads assumed to be turned up in tossing up a coin to a large no. of times is an example of a hypothetical population.

(ii) **Population member :** *Each of the individuals of mutually exclusive and exhaustive set of individuals comprising the whole population is called a member of the population.* For example, a field (or a plant) is a member of the population of fields (or plants) under investigation.

(iii) **Population size** : *The total number of members or sampling units contained in a population is called the population size.* Depending on the no. of members in the population, a population may be of two types—*finite* and *Infinite*. If a population contains only a finite no. of members in it, we call it a finite population. For example, the population of inhabitants of India, the population of books in a library, and the population of students in Meerut University etc. are the finite populations. On the other hand, a population with an infinite no. of members contained in it is characterized as an *infinite population*. For almost all practical purposes, a population consisting of a large no. of members may be considered as infinite. The population of pressures at various points in the atmosphere, and the population of throws of a die are the examples of an infinite population. In fact, the discussion of a finite population is more difficult than that of an infinite. Sometimes it is also difficult to ascertain whether a population is finite or infinite. The population of stars is an example of this type.

(iv) **Parameter** : *The population characteristic is called the parameter.* For example, the mean and the variance of a normal population are its two parameters.

(v) **Sample** : *A selected number of population members is called the sample.* It may also be termed as a representative part of the parent population which is selected for drawing conclusions regarding the population with a specified degree of confidence. In most of the problems, it is not practically feasible that every member of the population may be examined, but only a part of it. For example, if we are interested in finding the average yield of a particular crop in a certain locality, it is not possible rather advisable to take every field of that crop into consideration. Similarly, if one inquires into the average heights of the human population of India, he cannot afford the time and expenditure required to measure the height of each individual. In all such cases, an investigator examines only a limited number of individuals or units of the parent population and expects that these selected units truly represent the whole population.

(vi) **Sampling unit** : *A population member itself or an aggregate of it selected for the sample is called a sampling unit.* For example, either an individual or a household, or a village etc. may be taken as the sampling unit in a survey on human population.

(vii) **Sample size** : *The number of sampling units selected from a given population is called the sample size.* As a matter of

fact, the problem of choosing an appropriate size of the sample involves some technical difficulties. Since too large a sample size though assures more reliable results owing to the reduction in the sampling standard error of the estimate, but at the same time it raises the cost of survey and requires a considerable time and other resources. On the other hand, too short a sample size though reduces the cost, time and resources of survey, but the results obtained therefrom are less reliable since the sample is rarely a true representative of its parent population. Thus the middle course to be adopted to choose a moderate sample size is subject to several conditions of specified precision, cost and technique of sampling used. Sometimes the size is selected so as to give the minimum variance or the maximum precision of the estimate within the specified cost of survey, while in some other cases it is chosen to give the specified precision of the estimate at the minimum possible cost. Hence, the problem of determining an optimum sample size is undoubtedly a quite difficult task whose discussion is beyond the scope of this book.

(viii) Estimate : *The value of a population parameter computed from the sample is called the estimate.* For example, if a random sample of n observations in x is drawn from a normal population with mean μ and variance σ^2 , then the sample mean $\bar{x}(=\Sigma x/n)$ and the sample variance $s^2[=\Sigma(x-\bar{x})^2/(n-1)]$ are said to be the respective estimates of the population parameters μ and σ^2 . Further, an estimate is said to be the best if its average value is equal to the parameter value and its s.e. is minimum, i.e. a best estimate must satisfy the properties of unbiasedness and minimum variance. An estimate, sometimes, is also known as a *statistic*.

Thus an estimate or statistic is a suitably chosen function of sample observations. The value of this statistic is computed for a number of samples each of the same size and drawn from the same universe. But the values of the estimate thus obtained are usually different due to the fluctuations of random sampling. This series of different values under certain conditions follows some definite statistical frequency distribution. If the no. of samples drawn be larger and larger, this frequency distribution tends to a continuous distribution, most probably a normal one. This distribution of the sample estimate is called the *sampling distribution*. For an unbiased estimate, the mean of this distribution is always equal to the corresponding population parameter value. Also the s.d. of this distribution is called the s.e. of the estimate. If this error for a

unbiased estimate is minimum, then the chosen sample is supposed to be a true representative of its parent population.

(ix) Sampling error : *The error that occurs in a sample result is termed as the sampling error.* In fact, however great care we take in selecting an appropriate size of the sample and also the technique of sampling, some random sampling errors are inevitable in the sample results. The average magnitude of these random sampling errors known as *the s.e. of the estimate* depends upon the no. of units chosen in the sample (n), the original variability in the material of the population (σ^2), the sampling technique employed and the method of estimation used. It is obvious that the variability in the population material is beyond the control of an investigator. As regards the sample size it is known that the s.e. of an estimate is inversely proportional to the square root of the no. of units in the sample. Thus the amount of s.e. of an estimate can be minimized only by the refinement of the techniques of sampling and estimation, and also by choosing an appropriate sample size subject to the condition of time, cost and precision as desired.

(x) Sampling frame : *The description of the available information on all the sampling units in the population is called the sampling frame.* A sampling frame identifies the sampling units clearly and accurately.

(xi) Sampling fraction : *The ratio of the size of a sample to that of its parent population is called the sampling fraction or sampling ratio.* For example, if a sample of n units is selected from a population of N units, then the ratio n/N is termed as the sampling fraction. If a sampling fraction is very low, i.e. n is very small relative to N , then a quantity $(1 - n/N)$ or $(N - n)/N$ is called *the finite population correction factor (fpcf)*. In practice, a fpcf can be ignored whenever the sampling fraction does not exceed 5%, or even 10% in some of the problems. Thus if a sampling fraction is too low or a fpcf is close to unity, then the population size as such has no direct effect on the s.e. of an estimate. But in fact, the effect of ignoring a fpcf from the expression of the s.e. of an estimate is to overestimate the s.e. concerned. Also the inverse of a sampling fraction, i.e. the ratio N/n , is called the *raising factor* or the *expansion (inflation) factor*.

(xii) Sampling : *The process of selecting a sample from the given population is called the sampling.* In a true sense, the sampling must be an unbiased and representative one. It clearly means that

the method of selecting a sample should be such that an estimate furnished by the chosen sample must on an average be equal to its true parameter value with minimum standard error. In practice, a good no. of sampling techniques is available in the theory of sampling. The sampling technique can however be varied according to the nature of the population to be sampled. Some of the sampling techniques are—random sampling, stratified sampling, systematic sampling, cluster sampling, multistage sampling, multiphase sampling, quota sampling, purposive sampling, probability sampling, balanced sampling, systematic area sampling, sequential sampling, and ratio, regression sampling.

8.2 Purpose of sampling theory : The purpose of sampling theory is to make sampling more efficient so that we may get a reliable and as much information as possible regarding the parent population. It attempts to develop the procedures of sample selection and of estimation that provide, at the minimum possible (or fixed) cost, the estimates which are precise enough for the purpose of enquiry at hand. Thus, how far the sample represents the population and how to select a representative sample are the questions that the theory of sampling attempts to answer. The main principle adopted in the theory of sampling is *the logic induction* where we move from a particular situation to general. Knowing the form of the parent population, we usually intend to estimate the parameters or specify the limits within which the true parameters are expected to lie with a specified degree of certainty. It is, however, to be clearly understood that all sampling results are expressed in terms of probability.

8.3 Role of sampling theory : It attempts to get the sample estimates as precise as possible (or of specified precision) within the specified cost (or at the lowest possible cost) and limited resources at hand. The principle of specified precision at the minimum cost recurs repeatedly in the presentation of sampling theory. In any specific situation, we cannot account in advance the exact amount of precision since we do not know readily how large a sampling error will be present in the sample estimate. Instead, the precision of a sampling technique used is determined by the frequency distribution generated by the estimate provided the technique is repeatedly applied to the same population.

A considerable part of sampling theory deals with the computations of various formulae for determining the sample variances of the estimates that are obtained by employing the different

sampling techniques. For samples of moderate sizes that are common in practice, there is often good ground to assume that the sample estimates are approximately normally distributed. The sample variance of an estimate gives a basis for the measure of precision of the estimate in inverse terms, since the precision is inversely proportional to the variance of the estimate. The sampling theory applicable to sample surveys is quite recent and contains a good no. of new appreciable developments. The modern sampling theory is based on the finite populations whereas the older theory of sampling is confined to infinite populations. Further, the theory of sampling can be studied under two categories – (a) *the sampling of attributes*, and (b) *the sampling of variables*. The former is based on the qualitative data while the later is concerned with that of quantitative.

8.3 (a) The sampling of attributes : Here we are concerned only with the possession or non-possession of some specified qualitative characteristic (trait) or attribute by the units chosen in the sample. If a selected unit possesses the specified trait, it is known as the *success* while its non-possession of the trait is called the *failure*. The most of the sampling theory is based on the fundamental assumption of random sampling which may or may not be a simple. By simple sampling we mean the random sampling where each event has the same probability of success (p) and its materialization is independent of the successes or failures of the preceding trials. *Thus a simple sampling is necessarily a random sampling but a random sampling may not always be a simple sampling.* This concept may be clear from the following example of sampling with, or without replacement.

Suppose an urn contains 6 white and 4 black balls each of the same size and shape. Then the prob. of drawing a black ball at the first trial is $4/10$, while at the 2nd trial is $3/9$ provided the drawn ball is not replaced back to the urn. Hence obviously, the above two probabilities are different and the sampling though random is not simple. It means that if the previously drawn ball is replaced back to the urn before the next draw, then the procedure of drawing or sampling the ball is termed as *sampling with replacement*, while the contrary case is *sampling without replacement*.

Further we note that if we deal with a finite population, the random sampling may or may not be a simple sampling according to the aspect of with or without replacement. But a random sampling from an infinite population may always be treated as a simple

random sampling since the drawing of a unit does not materially affect the probability distribution. Also the selection of a simple sample of size n from a population of N units is equivalent to a series of n independent trials with constant probability (p) of success, or $q=(1-p)$ of failure. The probabilities of $x=0,1,2,\dots,n$ successes are the terms in the *binomial expansion* of $(q+p)^n$, which gives us the sampling distribution of the no. of successes in the sample. The mean of this distribution is np and the s.d. is \sqrt{npq} . Similarly, the mean of the proportion of successes is p and their s.d. is $\sqrt{(pq/n)}$. Hence the precision of the proportion of successes in the sample, being the reciprocal of their s.d. or s.e., is $\sqrt{(n/pq)}$ which is directly proportional to \sqrt{n} since p and q are constants.

8.3(b) The sampling of variables : Here we are concerned with the actual measurement on the units chosen in the sample for some specified quantitative characteristic or measurable value of a variable like the yield, or height, or weight etc. Each of the population members to be sampled gives its own measurement and all together they comprise some definite frequency distribution. A simple random sample of size n from a finite population of N units can be drawn in ${}^N C_n$ ways while an infinite no. of samples can be drawn from an infinite parent population.

Below we discuss in brief some of the techniques of selecting a sample from a given population.

8.4 Random sampling : *The method of selecting n units from a population consisting of N units is called random sampling (or strictly simple random sampling) if all ${}^N C_n$ possible samples have an equal chance of being chosen.* The sampling is based on the fundamental assumption that the population is homogeneous with respect to the character under study. In practice, a simple random sample is drawn unit by unit and any previously drawn unit is not replaced back so that it may not repeat in the sample more than once. This type of sampling is called random sampling without replacement. At any stage in the draw, this method gives an equal chance of selection to all the units not previously drawn. It should not, however, be forgotten that a random selection does not mean a haphazard selection. Also a random sampling from a finite population with replacement is equivalent to sampling from an infinite Population without replacement.

8.4.1 Method of selection : Certain methods of drawing a random sample that are convenient for small populations are not always suitable for large populations. The method of taking a

random sample depends to some extent on the size and nature of the population to be sampled. Thus the method of taking a random sample of 100 students from the population of students of a college may not be suitable in taking a sample of 1 kg. flour from a sack of flour containing 100 kgs. of flour. Two methods of selecting a random sample—(a) *sampling by lottery*, and (b) *sampling by random numbers*, are described below.

8.4.1(a) Sampling by lottery : In this method of sampling there corresponds a no; any of 1 to N , to every unit of the population to be sampled. These numbers are written on separate slips of paper of equal size and the lotolaty of these slips is mixed thoroughly in a bowl or estating dru.n. Finally a blind fold selection is made from the container, the selected no. of slips, say n , being equal to the size of the sample and the nos. on them being the population units to be selected. The practical difficulty of the method lies in shuffling or mixing the miniative population. If the population to be sampled is large enough, the constriction of the miniature population is also large. Thus the method can be suitable only in the case of small populations, since a thorough mixing of numbered slips is essential for complete elimination of bias.

8.4.1(b) Sampling by random numbers : To avoid some difficulties of lottery sampling, the random number tables are used to derive a sample from the specified population. Here as well every unit of the population is allotted a no; any of 1 to N . Instead of forming a miniative population we take any page of random sampling numbers and note the random numbers (n) equal to the size of the sample either in a horizontal or a vertical line (neglecting zero, repeated numbers and any no. greater than N). Though several persons have constructed the random no. tables but those of the following are commonly used.

(i) **Tables by Tippett :** The random number tables (Random Sampling Numbers, Tracts For Computers. No, 15, published by Cambridge University Press) constructed by L.H.C. Tippett are most satisfactory. These tables consist of 10400 four digit nos. which are constructed out of 41600 digits taken from census reports by combining them in fours. Although these nos. were chosen haphazard, yet their application in numerous investigations has shown their truthfulness.

(ii) **Tables by Fisher & Yates :** The random number tables (Statistical Tables For Biological, Agricultural and Medical Research workers, published by Oliver and Boyd) constructed by prof. R.A.

Fisher and F. Yates are of great importance in various statistical surveys.

(iii) **Tables by ISI, Calcutta** : The random number tables constructed by the authorities of Indian Statistical Institute, Calcutta are also very common in statistical investigations.

(iv) **Tables by Kendall & Smith** : The random number tables (Random Sampling Numbers, Tracts For Computers : No. 24, published by Cambridge University Press) constructed by prof. M.G. Kendall and B.B. Smith are very popular in numerous statistical surveys. These nos. are obtained by using a randomizing machine, and are quite reliable.

8.4.2 Assumptions of random sampling :

- (i) Each population unit has an equal chance of being chosen in the sample.
- (ii) Each possible sample has an equal chance of being chosen as a sample.
- (iii) The selection of the units is independent and free from any human bias.
- (iv) The population is homogeneous with regard to the character under study.

8.4.3 Properties of random sampling :

- (i) It is the easiest possible method of selecting a sample.
- (ii) It gives an equal chance to every population unit to be sampled.
- (iii) It does not require any extensive plan.
- (iv) It saves a considerable time, labour and cost of survey.
- (v) It gives a true representative of the population provided the sample size is suitably chosen.
- (vi) It gives the least chance to human bias.
- (vii) The precisions of its sample results can easily be examined.

8.4.4 Limitations of random sampling :

- (i) It cannot be applied to the situations where some population members are necessary to be selected in the sample.
- (ii) The method is not applicable to heterogeneous populations.
- (iii) It does not give the reliable results through small samples.

Exp.(1) Select a random sample of size 20 from a population of 8500 sticks.

Sol : Let all the sticks of the population be numbered from 1 to 8500 in some order. Now we consult a page of Tippet's random number tables and select the first 20 nos. either row-wise or column-wise such that none of them is zero, or repeated, or

greater than 8500. If the nos. obtained are : 3200, 3525, 3394, 1985, 7693, 0011, 3142, 4625, 7518, 2472, 5182, 7844, 7780, 3362, 3857, 6058, 4505, 1940, 7305 and 7947 then the sticks bearing these nos. would constitute the desired random sample of 20 sticks.

Note: Though Tippett's random nos. are four digit nos. but in many cases they are less than 1000 or even 10, e.g. 0058. Thus even if the units (N) in the population are less than 100, a sample of any size ($n < N$) can be drawn from these tables. In such cases some modified methods are available, but the discussion is omitted.

8.5 Stratified sampling : The method of selecting n units from a population consisting of N units is called stratified sampling (or strictly stratified random sampling) if k simple random samples of sizes n_i are independently drawn from k sub-populations of respective sizes N_i such that $n = \sum n_i$ and $N = \sum N_i$ for $i=1, 2, \dots, k$. These sub-populations are non-overlapping, homogeneous within themselves and are called *strata*. The sampling is based on the fundamental assumption that the population is markedly heterogeneous with respect to the character under study. For example, the human population of a country may be stratified according to the age-groups, or social circumstances, or economic conditions etc. Thus in stratified sampling, the population of N units is first divided into k distinct strata of N_1, N_2, \dots, N_k units and then simple random samples of n_1, n_2, \dots, n_k units respectively are drawn from these strata independently which all together comprise the whole sample so that $n = n_1 + n_2 + \dots + n_k$. The technique of determining the sample sizes of the strata, where n is known, is called the problem of *allocation*. Sometimes the sample-sizes are taken proportional to the stratum-sizes and sometimes proportional to the s.d.s. within the strata. Infact, the discussion of allocation is beyond the scope of this book.

Below we give some situations in which the technique can successfully be employed –

- (i) When the sampling problems are markedly different in different parts of the population.
- (ii) When several field offices are required for administrative convenience in supervising the survey work.
- (iii) When the results of specified precision are wanted for only a certain sub-division of the population.
- (iv) When some gain in precision in the estimates is desired over random sampling.

Exp (2), If a population of 1000 units is classified into four groups consisting of 100, 200, 300, and 400 units respectively with regard to some specified character. Outline the procedure of selecting a stratified sample of 20 units.

Sol : Let each of the four groups or strata be treated as a population in itself. Now using the method of random sampling, we can draw the simple samples of sizes 2, 4, 6 and 8 respectively from the strata of sizes 100, 200, 300 and 400. Thus the whole sample of size $n=(2+4+6+8)=20$ is the desired stratified sample from the given population.

8.6 Systematic sampling : *The method of selecting n units from a population consisting of $N(=nk)$ units is called systematic sampling if the first unit ($<k$) for the sample is selected at random from the first k units of the numbered population and then every k th unit thereafter.* The sampling is based on the assumption that the complete list of the population members is available. Thus in a systematic sampling of n out of N , the population units are serially numbered from 1 to N in some order and are such that $N=nk$, where n, k are both integers. Here we first draw a random number less than k , say i , from the first k units and then every k th subsequent unit in the population. Thus the systematic sample contains the n units : i th, $(i+k)$ th, $(i+2k)$ th, $\dots, [i+(n-1)k]$ th at regular spacings. For example, the selection of every k th block from a list of blocks, or the selection of every k th time interval for observing the no. of telephone calls on a public telephone-booth etc. can give us a systematic sample provided the first unit, less than k , is chosen with the help of the random no. tables.

A systematic sample is not truly random because of the fact that only the 1st unit for the sample is selected randomly which determines the whole sample, since the remaining units are automatically determined by a constant interval. But it may be equivalent to a simple random sample if the numbering of units in the population is effectively random. The alphabetical list of the names is a random list. It may also be equivalent to a *stratified random sample* provided the units in each stratum are randomly listed. Because in stratified sampling the unit to be chosen from each stratum is based on random selection while in a systematic sampling its position relative to the unit in the first stratum is readily determined. A systematic sample also resembles a *cluster sample* being equivalent to a sample of one cluster chosen out of the k clusters of n units each.

The systematic sampling has the following advantages—

(i) Drawing a sample is easier, less time consuming and often without mistake.

(ii) The systematic sampling is spread more evenly over the whole population and is thus sometimes more precise than stratified random sampling also.

Below we give some situations in which the method can successfully be employed—

- (i) When there is no periodicity in the list of the sampling units.
- (ii) When the k th units which constitute the sample are not alike or correlated.
- (iii) When we want simplicity and low expenditure on survey.
- (iv) When the complete list of population members is available.

Exp. (3) Obtain a systematic sample of 20 from a list comprising the population of 1000 individuals.

Sol : First we number the population members from 1 to 1000 in some order. Then taking $N=nk$ i.e. $1000=20k$, we get $k=50$. Thus we have to select the 1st unit for the sample between 1 and 50 from the random number tables. Let this no. selected at random be the 35th unit of the population. Thus the subsequent units in the sample are : (35+50)th, (35+2×50)th, ..., (35+19×50)th, i.e. 85th, 135th, ..., 985th. The sample, therefore, is constituted by the 20 units viz. 35th, 85th, ..., 985th.

EXERCISE VI.1

1. Select a random sample of size 10 from a population of 245 fields.

2. Explain the terms : sample, population, random selection.

(M. Sc. Ag, Agra, 1965)

3. Explain in brief the following—

random sampling, stratified sampling and systematic sampling.

4. Compare and contrast the merits and drawbacks of sample and census studies.

5. What is meant by sample-methods of enquiry? When it is adopted and what are its advantages?

6. Describe the special features of the different types of universes from which samples can be drawn.

7. Suggest a method of obtaining a random sample of words from the English language by the use of random sampling numbers and a dictionary.

8. Comment whether the following samples are representative.

(a) A mixture of sand and saw dust is sampled by taking a small quantity from the bottom.

(b) A basket of grapes is sampled by taking a handful from the top.

(c) Investigators into the size of the families in a town conducted a house-to-house inquiry in the after noon, ignoring those houses at which there is no reply.

9. Obtain a systematic sample of 10 from the list of voters comprising a population of 500 voters.

Part II

The Experimental Designs

Chapter I

Design of Experiments

Meaning & Definition :—

Men have always been learning by experience from the experimental observations. The observations obtained from a carefully planned and well designed experiment in advance give entirely valid inference. In fact, the inductive inference is the only method by which the new knowledge comes to this world. Any inference drawn from a sample regarding its parent population is always attended by some degree of uncertainty which may be defined by the method of Mathematical probability. With this assumption, *we define the Design of Experiment "as that logical construction of the experiment in which the degree of uncertainty with which the inference is drawn may be well defined."*

The principles of the design of experiments have so far been most explicitly developed in the field experimentation. In the field experimentation, we compare the different varieties of a crop, the different fertilizers, the different methods of seed-treatment and sometimes the different pieces of land itself. Thus, it is common to test the yield performance of a number of new varieties in comparison to a standard variety and also to examine the response of a crop to graded application of one or more fertilizer treatments. On the other hand, we may be interested in knowing the effect of the different cultivation processes. *These objects of comparison are called treatments.*

Suppose, we have a large homogeneous field divided into different plots and we apply different treatments to these plots. Then the yield of these different plots are recorded. If some of the treatments produce bigger effects than the others, it remains to the experimenter to decide whether the observed differences are due to the treatment effects or due to chance (uncontrolled) factor. Our past experiences tell us that the yields of the plots will vary even under the same treatment. This variation from plot to plot is due to the uncontrolled (chance, random) factors and is called the *experimental—error*. In order to test the significance of the difference between the two treatments, first we require an estimate of the experimental error and then apply the approximate test of significance. The basic -requirement for the former is to repeat the treatments to a number of times and for the later the random allocation of the treatments to various plots. We always desire a lower magnitude of the experimental error since the lower error detects the smaller real differences, i, e. it increases the precision of the design. It can be achieved partly by replications (repetitions of treatments) and mainly by adopting the *technique of local-control*. The local-control is the technique of dividing the whole experimental field which may be expected to be heterogeneous with regard to soil fertility in to homogeneous blocks row wise, column wise or both according to the fertility gradient present in the soil of the field. Thus the basic principles of the design of experiments are—

- (1) Replication,
- (2) Randomization and
- (3) Local control.

(i) **Replication** :—The repetition of the treatment under comparison is called the replication. The purpose of replication is two fold—

- (i) It reduces the experimental error. Since we know

that the sampling variance of a mean yield is $\frac{\sigma^2}{r}$. Where

' σ ' is the S. D. of the individual observations (i. e. the plot yield and 'r' is the number of replications. Therefore, the replication has an important but limited roll in increasing the precision of the design.

—(ii) The main purpose of replication is to supply an estimate of the experimental error of which there is no other alternative and without which the significance of the difference between the two treatments can not be judged.

(2) **Randomization** :—The allocation of treatments to various plots in a random manner is called the randomization. The purpose of randomization is:—

(i) To guaranttee the validity of the test of significance as the A. V. test (which is used to test the homogeneity of the data i. e. to test whether the different treatments are equally effective) is based on the assumption of the randomness of the observation.

(ii) To ensure that the different treatments on the average are subject to the same environmental effects. Therefore the difference between any two treatments remains free from bias.

Thus, when the treatments are replicated a number of times and allotted randomly to various plots in the field, we are in a position to test the significance of the observed treatment differences by the help of statistical tests.

(3) **Local Control**:—When the experimental area is heterogeneous and the treatments are scattered randomly over the whole area, the soil heterogeneity will also enter the chance factor and thus increases the experimental error. It is desirable to reduce the experimental error as far as practicable, since the lower experimental error can detect a smaller real difference between the treatments. In order to remove the soil fertility effect from the experimental error; the whole experimental area is divided into homogeneous groups (blocks) row wise, column wise or both according to fertility gradient in such a way that the variation between the blocks is maximum

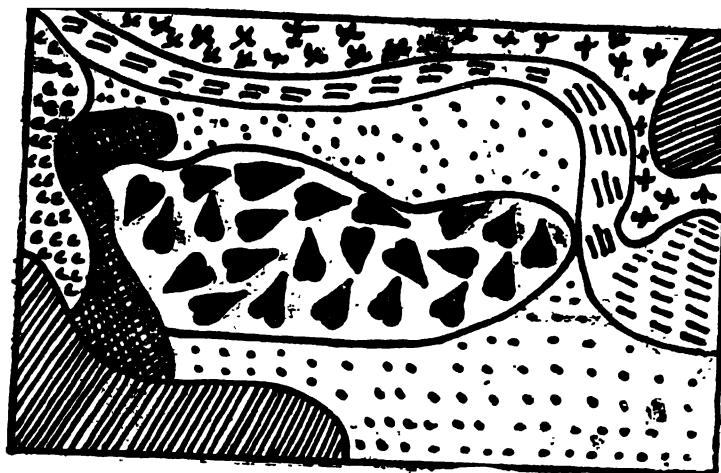
and within the blocks is minimum. The randomization is kept restricted over the blocks. This process of reducing the experimental error by dividing the experimental area into more homogeneous blocks is known as *local control*. The introduction of local control ensures that the comparison between the treatments is made under as similar conditions as is possible with resources at hand and this helps in the reduction of the experimental error. Various forms of arranging the plots into homogeneous blocks have so far been evolved which are called *experimental design*.

Uniformity Trial:—As mentioned above that for reducing the experimental error we have to divide the whole experimental area into more homogeneous blocks, for that we must have a correct idea of the fertility variation in the field. This idea may be obtained from the result of uniformity trial. *A uniformity trial consists of growing the same crop with the same treatment all over the field.* The whole field is divided into several small plots of equal size and the yield from each of these units (plots) are recorded separately. From such records, we can prepare a fertility contour map which gives a good idea of the nature of the soil fertility variation. The fertility contour map is prepared by joining the points of equal fertility through lines.

Fertility Contour map

An eye inspection of the fertility contour map shows that the fertility does not increase or decrease in any systematic pattern but its distribution over the whole field is random. It is also observed that adjacent units are more or less similar in fertility than those apart. A homogeneous block can be formed by combining a number of adjacent units. The number of such units, which will form a block, can be determined by calculating the coefficients of variation for several combinations of several units and choosing that combination for which the coefficient of variation (c. v.) is minimum. The variation in the plot-yields under uniformity trials is due to the

Fertility contour map



Different shades showing different fertility

uncontrolled factors called experimental error. Thus a uniformity trial gives an estimate of the experimental error.

In short, an uniformity trial gives—

- (i) an estimate of the nature and extent of the fertility variation.
- (ii) an estimate of the experimental-error, and
- (iii) a clue to reduce the experimental error by forming the homogeneous blocks.

Precision: The precision of the design is the ability with which it detects the smaller real differences between the treatments. The precision of a design is more than the other if the *least significant difference* (critical difference) between the treatments at a given level is lower in it than that of the other. In other words the prob. of observing a difference less than or equal to a given value measures the precision of the design. Thus the degree of uncertainty with which we draw our conclusion is called the precision of the experiment. To

find a more suitable design for a particular problem means to achieve the maximum precision with the given cost and resources. The precision of a design can be increased by decreasing the experimental error (random error). The lesser the experimental error, the greater is the precision. Thus the precision can be increased by increasing the no. of replications and using the technique of local control.

Accuracy: In any experiment, the plot-yields are also affected by a no. of other factors than the treatments such as cultural operations (ploughing, hoeing, weeding, earthing etc), manurial doses, cultivation processes etc. If the effects of these factors are not the same on various plots, then the treatment differences will be subject to a constant bias (systematic error) which cannot be diminished by increasing the no. of replications like the experimental error. In order to diminish this bias, the experimental technique should be so refined that all the plots are equally affected by the above factors. The lesser the amount of bias the greater is the accuracy of the design. Thus the accuracy of a design is a measure for the lack of bias.

Experimental Material: The material, on which the experiment is performed, is called the experimental material e. g. agricultural field, herd of cows, patients in a hospital and plants in a green house etc.

Experimental unit: The whole experimental material is divided into a no. of small parts to which the treatments are applied. These small parts are called the experimental units e. g. plot of a field, cow in a herd and plant in a green house are the experimental-units.

Following are the main considerations in planning an experimental design—

(i) Object : (Formulation of hypothesis to be tested) : Every experiment has a definite object to achieve and this object is indirectly defined by the null hypothesis (H_0). The null hypothesis (or object) must be clearly and exactly stated without any ambiguity. For example, if we want to compare the effects of a number of manurial treatments on the yield of a certain crop, then we must decide whether the yield is a grain yield, straw yield or total produce (grain yield + straw yield). In addition to make a decision regarding the yield, we must also decide the variety, irrigation conditions, nature of the soil and cultural treatments etc. to be used in the experiment. In absence of the knowledge of all such relevant details, neither valid conclusions can be drawn from the experimental-data nor the scope of the experiment can be decided. Therefore, a clearly defined object is an essential part in the planning of an experiment.

(ii) Scope : The conditions under which the experimental results are valid, decide the scope of the experiment. For example, if the results of a design are valid for a particular variety of a crop and soil, while that of other are valid for any variety and soil then the former is of limited scope than the later. The scope of an experiment can be widen by testing a number of factors and their levels simultaneously. It is desirable to have a sufficient scope of the experiment as far as the experimental material and the cost permit.

(iii) Feeler experiment : Suppose, the object of the design is to compare a foreign imported variety of a certain crop against a local variety. Then, before starting the actual experiment it is necessary first to know whether the imported variety will germinate and prove itself a successful variety under the changed climatic and soil conditions. This can be known by showing the new variety in some plots, called the *observation plots*. Such an experiment which is carried out to test the suitability of some treatments is called *feeler*

experiment It is recommended for the situations where the suitability of some treatments is doubtful otherwise it will be a waste of resources and time.

(iv) Experimental Site:—

The experimental site should be as homogenous as possible. The idea of uniformity can be had by having a glance of standing crop, surface or better from the uniformity trial data. The uniformity trial is recommended for the newly acquired lands for which we do not have a pre-idea of fertility variation otherwise it will delay the experiment. In the field experimentation, it is very difficult to have a uniform experimental site, the fertility gradient will be present in one or more directions. Another care in selecting the experimental site is that there should be no tree on its border as the shade of the tree affects the yields of the border-plots. In the case, there is a tree on the border, the area which is expected to be affected by the tree should be excluded from the experimental area.

(v) Choice of the experimental design:—

The choice of the experimental design depends upon the heterogeneity of the experimental site, no. of treatments and the relative precision with which the treatments are to be compared. If—

(1) the experimental site and environment are uniform, then a C. R. D. is used. This design compares all the treatments with equal precision.

(2) the experimental site is not uniform but can be grouped (according to a single criterion of classification) in to homogeneous groups (blocks) of land then R. C. B. D or R. B. D. is used provided the no. of treatments is not large otherwise *Incomplete Block design* will be used.

(3) the experimental site is not uniform but can be grouped according to a double criteria of classification in to homogeneous groups of land, then L. S. D. is used provided the no. of treatments ranges from 5 to 12. In the field

experimentation, this situation arises when the fertility gradient is in two directions at right angles.

(4) some of the treatments are to be compared with relatively higher precision than others, then the Confounding Scheme is used provided the precision of the higher order interaction is to be sacrificed.

(5) some of the treatments require larger plots and are to be studied with relatively lower precision than the others, the S. P. D. is used.

Another consideration for the choice of the experimental design is the availability of the resources. For example, the lack of proper training and Skill prevent the use of complex design even if they are of higher precision.

(vi) **Replication:** An adequate no. of replications for a no. of treatments cannot be suggested in advance of the experiment as it requires the knowledge of the fertility-variation in the experimental site which is rarely known. In absence of this knowledge, the rule for the no. of replications is to take such a no. of replications that provides at least 12 d. f. for error. This rule is based on the fact that the values of F do not decrease rapidly beyond $v_2 = 12$.

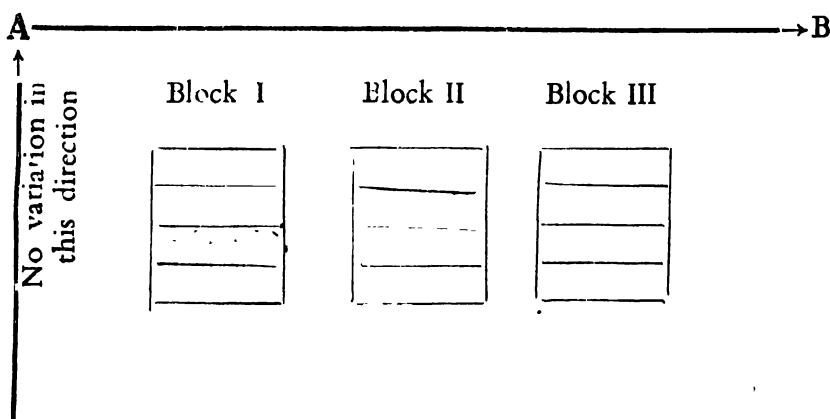
(vii) **Randomization:** For the validity of the experimental-results, it is necessary that the different treatments should be randomly allocated to different experimental units (plots). The procedures of randomization are different for different experimental designs.

(viii) **Refinement of experimental technique:** In order to achieve the real treatment differences, it is essential that all the experimental plots should be subjected to the same type of cultural operations other than those under investigation.

(ix) **Ancillary observations:** When the plot-yields are suspected to be affected by a character which is uncontrolled and varies randomly from plot to plot, it should be measured in addition to the yield of each plot. The measurements made on the uncontrolled character are called *ancillary observations*.

These observations are used to eliminate the variation due to the uncontrolled character from the experimental error and thus to increase the precision of the experiment.

(x) **Shape of blocks and plots:** When the experimenter has a choice for the experimental area, he should choose that one which seems to be uniform. An idea of the uniformity can be obtained from the appearance of the surface and the previous crop. After the selection of the experimental area, he has to investigate the fertility gradient of the area. Suppose, the fertility-gradient is in the direction from A to B.



Now he has to divide the whole experimental area into different homogeneous blocks such that the variation between them is maximum. This can be achieved by making the blocks as compact (square) as possible. These compact blocks are arranged one after one along the fertility-gradient. Further, each block will be divided into as many plots as the no. of treatments. In dividing the blocks into plots, the object is just opposite to the above i. e. the variation within the blocks should

be minimum. This object can be attained when the plots are as alike as possible to each other. Thus the shape of the plots should be rectangular with its larger side parallel to the direction of the fertility-gradient and all of them should be in a row across the block as shown in the above figure. When the no. of treatments is large, the plots have to be arranged in two or more rows in order to maintain the compactness of the blocks. An exception of this, is the case of the sloping experimental site. In this case, the plots are arranged in a single row with their longer sides parallel to the slope as the fertility-gradient is in the direction of the slope. This arrangement sacrifices the compactness of the blocks in the case of larger no. of treatments.

(xi) **Size of the plot** : The size of the plot depends upon the experimental area available, no. of treatments and their replications and the crop. The optimum sizes of the plots for different crops are given in the following table.—

S.N.	Name of the crop	Plot-size in acre
1.	<u>Cereals</u>	1/10
2.	<u>Maize</u>	1/20
3.	<u>Sugar-cane</u>	1/40 to 1/20
4.	<u>Vegetable</u>	1/80

(xii) **Border effect** : The borders of a plot are also affected by the treatment given to the neighbouring plots while the central plants remain unaffected. In order to eliminate this effect, a non experimental border should be left around each plot which is given the same treatment and cultivated in the same way (cultivation practices) as the plot, but its yield is harvested separately and is not taken into consideration for the purpose of the experiment.

(xiii) **Statistical Analysis**: The consideration discussed so far, enable us to draw the valid conclusions of high precision while the statistical analysis provided a way how the conclusions can be drawn from the recorded experimental data. The statistical analysis comprises of—

(i) analysis of variance,

(ii) computation of S.E. & C.D.,

(iii) a sketch of the tabular form for presenting the results and

(iv) an account of the test of significance to be applied to the proposed design.

(xiv) **Report** : Finally, the conclusions drawn from the statistical analysis and comments on them (if any) should be summarized in the form of a report.

EXERCISE NO. I

Q. No. 1 :—Explain the terms given below, mentioning their rolls in the field experimentation:—

- (a) Local control,
- (b) Replication, (M. Sc. Ag. Agra, 1964)
- (c) Randomization and
- (d) Uniformity trials, (M. Sc. Ag. Agra 1958)

Q. No. 2 : Discuss the important practical considerations in carrying out field-experiments at the research farms?
(M. Sc. Ag. Agra, 1963)

Q. No. 3 : Define the following terms with a suitable example for each.

- (a) Experimental-unit,
- (b) Treatment,
- (c) Precision & accuracy and
- (d) Random (experimental) error.

Q. No 4 : Describe the important methods for increasing the accuracy of field experiments ?
(M. Sc. Ag. Agra, 1965)

Hint:—Replication, local control, ancillary observation, and refinement of experimental technique are the important methods.

Q. No. 5 : What are uniformity trials ? How do they help in determining the optimum size and shape of the experimental plots ?

(M. Sc. Ag. Agra, 1959)

Chapter II

Completely Randomized Design (C. R. D.)

Description:—

On the assumption that the whole agricultural field under experiment is homogeneous, we lead to the simplest type of design called the C. R. D. Though this assumption of homogeneity is too big and particularly it is never satisfied in agricultural field experimentation at least, still in many laboratory experiments e. g. in Physics, Chemistry and cookery where a quantity of material after thorough mixing, is divided into small samples (units) to which the treatments are applied, this procedure is the best and the first.

In this stage of layout, each treatment is allotted to different units entirely by chance (randomly). Particularly, if a treatment is to be applied to four units, the randomization gives every group of four units an equal chance of receiving the treatment.

Advantages:—

(i) All the experimental-material can be utilized and any number of treatments with different replications can be used.

(ii) The statistical analysis is easy even if the no. of replications are different for different treatments or if the experimental error differs from treatment to treatment.

(iii) The analysis remains simple even in the case if one or more units are missing or rejected. Moreover, the relative

loss of information $\left(\frac{-1}{\sigma^2}\right)$ due to missing data is smaller in comparison to any other design.

(iv) It provides the max. no. of d.f. to estimate the error variance.

Disadvantages:—

(i) The main demerit lies in the assumption of homogeneity. Suppose the whole experimental material is not homogeneous, then the whole variation among the units enters the experimental error, thus increasing the error variation and consequently making the design inefficient. Though for a given no. of treatments and experimental units, this design provides a max. no. of d.f. for the estimation of error and thus increases the sensitiveness of the experiment.

(ii) Due to the assumption of homogeneous and scarce material, this design is seldom used in field experiments and is replaced by a better substitute "Randomized Complete Block Design" (R.C.B.D.).

Applications:—

The C. R. D. is appropriate under the following situations—

(i) When the experimental material is homogeneous and limited as in laboratory experiments.

(ii) Where an appreciable fraction of units is likely to be destroyed or to fail to respond.

(iii) In small experiments, where the increased accuracy from an alternative design does not compensate the loss of error d.f.

Randomization:—

Suppose, we have got three treatments A, B, & C each replicated four times, so we divide our experimental-area into 12 equal sized-plots. Let these plots be numbered in a convenient way from 1 to 12. Then we consult a random number table and write down serially the numbers in order they occur

in the table neglecting zero, repeated numbers and those greater than 12. Let these numbers be—

1, 3, 6, 12, 4, 7, 9, 10, 2, 5, 8, 11. Now the first treatment A will be applied to the plots bearing the numbers 1, 3, 6 & 12, treatment B to the plots bearing the numbers 4, 7, 9, & 10 and treatment C to the remaining plots bearing the numbers 2, 5, 8 & 11.

Formulation of the hypothesis to be tested:

In order to test the significance of the difference between the treatment-means, let us set up the hypothesis—

H₀: that there is no significant difference between the treatment-means.

Then, we can calculate the prob. for the observed difference assuming the null hyp. (H_0) to be true. If we are unable to calculate this prob., we can not draw the definite conclusion from the experiment. This can only be done, when the null hyp. (H_0) is clearly defined. Therefore, the setting up of the null hypothesis is as much an essential part of the design for the interpretations of the results as replication and randomization are for the sensitiveness and validity of the experiment.

We must note that a null hyp. can be rejected and never be accepted. Every experiment is designed and performed in such a way as to give the max. chance for the rejection of the null hypothesis whenever it is wrong. As soon as the null hypothesis is rejected, we arrive at the definite conclusion. But if it is not rejected we say that there is no evidence against the null hyp. on the basis of the observations made.

Statistical Analysis:

Suppose, we have got 'v' treatments (t_1, t_2, \dots, t_v) replicated r_1, r_2, \dots, r_v times respectively, the plot-yields can be arranged in the following tabular-form—

Treatment	Plot-yield			Totals
t_1	y_{11}	$y_{12} \dots \dots y_{1r_1}$		T_1
t_2	y_{21}	$y_{22} \dots \dots y_{2r_2}$		T_2
\vdots	\vdots	\vdots		\vdots
t_v	y_{v1}	y_{v2}	y_{vr_v}	T_v
Totals	—			G

To break up the T.S.S. into two parts—

- (i) between the treatments (S.S. due to treatments) and
(ii) within the treatments, (S.S. due to error), we proceed in the following manner—

$$(i) \text{ T. S. S. } = \sum_{i,j} y_{ij}^2 - \text{C. F.}, \quad \text{where C. F.} = \frac{G^2}{N}$$

$$= S \text{ (Say)} \quad \text{and } N = \sum_i r_i, i = 1, 2, \dots, v$$

$$j = 1, 2, \dots, r_i$$

- (ii) S. S. due to treatments

$$= \sum_i \frac{T_i^2}{r_i} - \text{C. F.}$$

$$= S_1 \text{ (Say)}$$

- (iii) S.S. due to error = T.S.S. — S.S. due to treatments = S_2
(say)

The results are summarized in the following A.V.T. —

Source of variation	D. F.	S. S.	M. S.	cal. F	F at 5%	F at 1%
Treatment	$v-1$ $=v_1$	S_1	$\frac{S_1}{v_1} = V_T$	$\frac{V_T}{V_E}$	---	---
Error	$N-v$ $=v_2$	S_2	$\frac{S_2}{v_2} = V_E$	if $V_T > V_E$		
Totals	$N-1$	S	—	—	—	—

Note (1) : If the error variance (V_E) is greater than the variance of the factor of classification then F is calculated by keeping V_E in the numerator.

Note (2) : In the analysis of variance table, the calculated values of F significant at 5% level are marked with one star (*) and those significant at 1% are marked with double stars (**) and they are said to be highly significant.

Inference: According to the null hypothesis our treatments are equally effective (identical in their yielding capacity) and so under the hypothesis the variations between the treatments & within the treatments; both are due to chance causes and so they must not be significantly different.

under the above consideration,

if $F_{cal.} = \frac{V_T}{V_E}$ comes out to be greater than $F_\alpha (v_1, v_2)$, then

the hypothesis is rejected. (where α is the desired level of significance)

(ii) If $F_{cal.} < F_\alpha (v_1, v_2)$, there is no evidence against the null hypothesis at $\alpha\%$ level of significance.

There is one of the two outcomes of the above test—

(i) either the hypothesis is not rejected.

or (ii) the hyp. is rejected.

If the hyp. is not rejected, it means that over all there are no significant differences between the treatments and we do not require any further analysis. But if the hyp. is rejected, we conclude that the treatments have their significant effect and further want to know which of them is more effective. For

this purpose we shall compare the treatment-means in pairs by Student's 't' test given by—

$$t = \frac{[\text{treatment mean-difference}]}{\text{S.E. of the difference}}$$

$$\begin{aligned} \text{where, S.E. of the difference} &= \sqrt{V_E \left(\frac{1}{r_1} + \frac{1}{r_2} \right)} \\ &= \sqrt{\frac{2 V_E}{r}} \quad \text{if } r_1 = r_2 \end{aligned}$$

Critical-difference:

Instead of calculating Student's 't' for different pairs of treatment-means, we can find the least significant difference at a given level of probability. This difference is known as the critical-difference (C.D.) and is given by the formula—

$$(\text{C.D.})_{\alpha\%} = (\text{S.E. of the difference}) \times t_{\alpha\%} (\text{error d. f.})$$

where $t_{\alpha\%} (\text{error d. f.})$ stands for the tabulated value of t at $\alpha\%$ level for the error d.f.

If the difference between the two treatment-means comes out to be greater or equal to C. D. , they are significantly different otherwise they are insignificant.

While comparing the several treatment-means, we arrange them in the descending order of their magnitudes and then compare in pairs. The means which do not differ significantly are *under lined by a bar*.

Report:

The results obtained and the comments on them are summarized in the form of a report, in the end.

Exp. No(1):

Three treatments A,B & C are compared in a completely randomized design (C. R. D.) with six replications for each. The lay out and straw-yield in Kgm/plot are given in the following table—

A 17	B (19	A 29	C 33	B 23	B 21
B 15	A 25	A 17	C 35	C 29	B 23
A 33	C 25	B 19	C 37	A 23	C 27

Where A denotes the control.

B " " application of nitrogen
22 Kgm/ Hectre

C " " application of nitrogen
44 Kgm/ Hect.

Analyse the experimental - yields and state your conclusions ?

Solution:

For convenience of the statistical analysis, the data is arranged in the following *tabular-form*

Replication Treat.	I	II	III	IV	V	VI	Total = T	Means	(T) ²
A	17 (289)	29 (841)	25 (625)	17 (289)	33 (1089)	23 (529)	T _A = 144 (3662)	24	20736
B	19 (361)	23 (529)	21 (441)	15 (225)	23 (529)	19 (361)	T _B = 120 (2446)	20	14400
C	33 (1089)	35 (1225)	29 (841)	25 (625)	37 (1369)	27 (729)	T _C = 186 (5878)	31	34596
The values in the brackets denote the squares of the respective yields of the plots.								G = 450 (11986)	25 202500

∴ The treatments A, B & C do not differ significantly in their yielding-capacity.

$$C. F. = \frac{G^2}{N} = \frac{(450)^2}{18} = 11250$$

$$\begin{aligned} T. S. S. &= \sum_i \sum_j y_{ij}^2 - C. F. \\ &= (17)^2 + (29)^2 + \dots + (27)^2 - 11250 \\ &= 11986 - 11250 = 736 \end{aligned}$$

where y_{ij} denotes the yield of j^{th} plot under i^{th} treatment,

$$\left. \begin{array}{l} i = 1, 2, \dots, v \text{ \& } v = 3 \\ j = 1, 2, \dots, r \text{ \& } r = 6 \end{array} \right\} \begin{array}{l} \text{Also } N = \sum_i r_j \\ \quad \quad \quad = 18 \end{array}$$

S.S. due to treatments

$$\begin{aligned} &= \frac{T_A^2 + T_B^2 + T_C^2}{6} - C.F. \\ &= \frac{(144)^2 + (120)^2 + (186)^2}{6} - 11250 = 11622 - 11250 = 372 \end{aligned}$$

$$\begin{aligned} S. S. \text{ due to error} &= T. S. S. - S. S. \text{ due to treatments} \\ &= 736 - 372 = 364 \end{aligned}$$

The results are summarized in the following A.V.T.—

source of variation	D.F.	S.S.	M.S.S.	F cal.	F tabulated at	
					5%	1%
Treat.	2	372	186.0	7.658 ^{***}	3.69	6.37
Error	15	364	24.27			
Totals	17	736	—	—	—	—

The value of F cal. comes out to be highly significant, hence the three treatments differ significantly in their yielding capacity.

Further to see which of the treatment is more effective, we compute the S. E. of the difference between the treatment means and the C. D. in the following manner—

$$\text{S. E. of difference} = \sqrt{\frac{2\overline{V_E}}{r}} = \sqrt{\frac{2 \times 24.27}{6}} = 2.84$$

$$(\text{C. D.})_{.05} = (\text{S. E. of difference}) \times t(15)_{.05}$$

$$= 2.84 \times 2.131 = 5.9520$$

Now we write the treatment means in the descending order of their magnitudes—

Treat. C	A	B	
mean. 31	<u>24</u>	<u>20.</u>	The treatment means which do not differ significantly are under lined by a bar.

Report:

The three treatments A, B & C differ significantly in their mean yields and the treatment C has yielded more than A or B. The differences between the mean of C and that of B or A are significant. A has also shown more average yield than that of B but their difference is not significant. Thus we conclude that the application of nitrogen has increased the straw-yield. This is due to the fact that in the presence of nitrogen vegetative growth takes place.

Note:

In the above example, if we take the deviations from $y=25$, the calculations will become much more easy.

Exercise I**Qⁿ. No. (1):**

The data represent sugar-yield in tons/acre for five varieties of sugar beet.

Variety	plot-yields					
A :	1.33	1.35	1.35	1.39	1.38	1.40
B :	1.31	1.38	1.36	1.37	—	—
C :	1.35	1.32	1.34	1.31	1.36	1.33
D :	1.34	1.32	1.36	1.34	1.35	—
E :	1.33	1.35	1.33	1.36	—	—

Analyse the data and test for a significant difference between the varieties ?

Ans: $F=1.80$

Qⁿ. No. (2);

Three varieties A, B & C of a crop are tested in a completely randomized design with three replications for each. The layout and the yield in pounds/plot are given below—

A 8	B 20	C 14
B 22	C 23	A 18
A 7	C 11	B 18

Analyse the experimental-yield and state your conclusions ?

Ans : $F=2.5$

Qⁿ. No. (3) :

Five plants were selected from each of half a dozen varieties of pea and their pods were counted—

Varieties	Pods/plant				
V ₁	17	23	27	25	20
V ₂	6	9	7	6	4
V ₃	9	12	13	11	11
V ₄	14	7	17	20	17
V ₅	25	23	20	32	27
V ₆	67	59	53	61	72

Prepare the analysis of variance table to test the significance of difference between the average number of pods of the six varieties ?

Ans: $F=101.6$

Qⁿ. No. (4):

The following table gives the butter-fat percentage in cow's-milk for different four breeds—

Breed	Butter fat %					
A :	4.0	4.5	4.0	4.5	5.0	5.0
B :	5.5	5.0	4.0	5.0	4.5	4.5
C :	5.0	4.5	4.5	5.5	5.0	4.5
D :	5.5	6.0	5.0	6.5	5.5	4.5

Analyse the data and state your conclusions ?

Qⁿ. No. (5):

Ans: $F=3.84$

In a varietal trial involving six varieties of pea each with 4 replicates, the yields in lbs/plot are given below. Analyse the data and arrange the varieties according to their performance assuming the data to be homogeneous with respect to the replications ?

Variety	Yield in lbs/plot			
V ₁ :	17.8,	17.3,	28.5,	18.5
V ₂ :	20.6,	18.8,	29.5,	21.0
V ₃ :	17.7,	12.7,	26.8,	24.9
V ₄ :	6.2,	5.0,	9.6,	4.1
V ₅ :	6.2,	7.0,	5.4,	7.7
V ₆ :	14.9,	12.5,	16.3,	12.6

Ans : $F=12.2$

Qⁿ. No. (6)

[a] What is a completely randomized design ? Give its applications and advantages ?

[b] Following table gives the life-periods in weeks for 4 batches of radio-valves. Test the significance of difference between the life-periods of the four batches?—

Batch	Life in weeks				
A :	130	138	134	142	150
B :	138	134	146	150	142
C :	126	138	142	130	146
D :	118	106	122	114	118

Ans: $F=13.96$

CHAPTER III

Randomized Block Design (R.B.D.)

Description :—If the whole experimental area is not homogeneous and the fertility gradient is in one direction only, then it is possible to divide the whole area into the homogenous blocks perpendicular to the direction of fertility gradient. Each of the blocks constitutes a single replication. If the treatments are randomized within each block separately, the result is a *randomized block-design*. This design controls one source of variation in the experimental material. Since all the treatments will be applied in each block, therefore the blocks will be divided into as many plots as the number of treatments.

As the experimental error has to be estimated from variations within blocks, it is essential that the blocks should be as homogeneous as possible. We know that the total variability of the whole experimental-material can be divided into two parts :—

(i) Between blocks and

(ii) Within blocks (error). If we reduce one, the second will automatically increase since the total variation over the whole material is constant. Thus, we arrive at the conclusion that the variation between the blocks should be maximum and that of within blocks minimum. At this stage, we must also note apart from the treatments to be compared, the whole block (all the plots of a block) should have uniform agricultural treatment (weeding, hoeing, earthing, harvesting etc.). For example, if the hoeing is to be spread over a number of days then it must be done for all the plots of a block on the same day.

Applications :—In field experiments, this design is used when the fertility gradient is in one direction only. In fact the design can be used where it is desired to control one source of variation in the experimental material. For example, in comparing the effects of different diets on the milk-yield of cows when they are of K

different breeds or lactation periods, in comparing the effects of different drugs in controlling a certain disease when the patients are of different age-groups, in comparing the efficiency of a number of salesmen when they are sent to different types of sale-areas.

Merits & Demerits :—Following are the chief advantages of a R.B.D.

(i) **Sensitiveness** :—This design remove the variation between the blocks from that of within blocks which generally results in a decrease of experimental-error and thus increases the sensitiveness. 'Cochran' has shown that the experimental-error of R.B.D. is 60% that of C.R.D.

(ii) **Flexibility** :—This design allow any numbers of treatments and replications. The only restriction is that the number of replications are proportional to the number of blocks. Generally, the no. of replications is equal to the no. of blocks and if any how extra replication is desired to some treatments, they may be applied to 2 or more units within each block.

Although any no. of replications can be used with any no. of treatments but it is desired that the no. of treatments should be such that they provide at least 12 d. f. for error. If 'v' is the no. of treatments, then the min. no. of replications ensuring at least 12 d. f. for error is $r = 1 + \frac{12}{v-1}$

(iii) **Ease of analysis** :—The statistical analysis is simple even in the case of missing values.

(iv) **Unbiased comparisons** can still be made when the experimental-error-variance is different for different treatments.

No design is more popular than R.B.D. due to its sensitiveness, flexibility and ease of analysis.

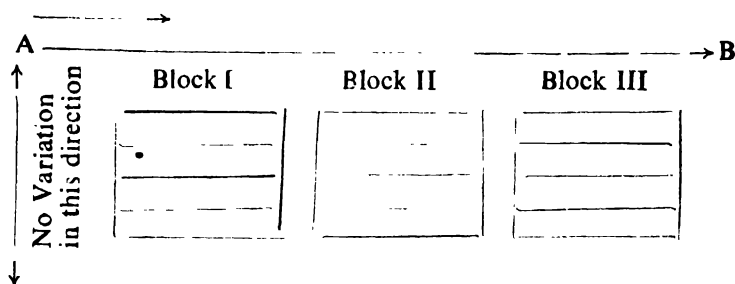
The demerits of this design lies in the fact that it cannot control variations in the experimental-material from two sources and in such a case it is not an efficient design. Further, if the no. of treatments is large then the size of the block will increase which usually results in introducing heterogeneity within block and thus increasing the experimental-error.

The above considerations lead that R.B.D. is not a suitable design when the no. of treatments is large and the variations in the experimental material are from two sources.

Shape of Blocks and Plots:—When the experimenter has a choice for the experimental area, he should choose that one which

seems to be uniform. An idea of the uniformity can be obtained from the appearance of the surface and the previous crop. After the selection of the experimental area, he has to investigate the fertility-gradient of the area. Suppose, the fertility-gradient is in the direction.

from A to B



Now, he has to divide the whole experimental area into different homogeneous blocks such that the variation between them is maximum. This can be achieved by making the blocks as compact (square) as possible. These compact blocks are arranged one after one along the fertility-gradient. Further, each block will be divided into as many plots as the no. of treatments. In dividing the blocks into plots, the object is just opposite to the above i. e. the variation within the blocks should be minimum. This object can be attained when the plots are as alike as possible to each other. Thus, the shape of the plots should be rectangular with its longer side parallel to the direction of the fertility-gradient and all of them should be in a row across the block as shown in the above-figure. When the no. of treatments is large, the plots have to be arranged into two or more rows in order to maintain the compactness of the blocks. An exception of this is the case of the sloping experimental sight. In this case, the plots are arranged in a single row with their longer sides parallel to the slope as the fertility-gradient is in the direction of the slope. This arrangement sacrifices the compactness of the blocks in the case of large no. of treatments.

Layout and Analysis

Randomization:—Suppose, we have got five varieties A, B, C, D & E and want to replicate six times each. Then the whole field must be divided into six homogeneous blocks each having five plots of equal size. The randomization is done by consulting a *Random Number Table* and selecting the one digit numbers in order they occur in the table leaving 0, repeated numbers and numbers > 5. Let these be—

1, 5, 2, 4, and 3. Now, we may take up first block, the treatment A will be given to 1st plot of this block, treatment B to 5th, C to 2nd, D to 4th and E to 3rd plot of the block. Thus the first block has received 5 treatments in a random manner. In each block, fresh randomization is made. Thus, six independent randomizations are needed for six blocks.

Formulation of Null Hypothesis:—In order to test the significance between the treatment means, we set up the null hypothesis that there is no significant difference between the treatment means and blocks.

Statistical Analysis:—Suppose, we have got 'v' treatments each with 'r' replications and if y_{ij} denotes the yield of the plot which is in the j th block and to whom i th treatment is applied. The yield from the original layout can be arranged in the following tabular form—

Treat/Block	1	2	j r	Totals
1.	y_{11}	y_{12}	y_{1j} y_{1r}	T_1
2.	y_{21}	y_{22}	y_{2j} y_{2r}	T_2
i .	y_{i1}	y_{i2}	y_{ij} y_{ir}	T_i
...
v	y_{v1}	y_{v2}	y_{vj} y_{vr}	T_v
Total	B_1	B_2	B_j B_r	G

If the original yields are large numbers, then their deviations from some arbitrary origin can be used in the above table.

Here, the sources of variation are—

- (i) treatments,
- (ii) blocks and
- (iii) Error.

The sum of squares are obtained as follows—

$$(i) \text{ T. S. } S = \sum_i \sum_j y_{ij}^2 - C. F., \text{ where } C. F. = \frac{G^2}{N} \text{ and } N = v r$$

$$= S \text{ (say)}$$

$$(ii) \text{ S. S. due to treat.} = \sum_i \frac{T_i^2}{r} - C. F., = S_1 \text{ (say)}$$

$$(iii) \text{ S. S. due to blocks} = \sum_j \frac{B_j^2}{v} - C. F., = S_2 \text{ (say)}$$

$$(iv) \text{ S. S. due to error} = T. S. S - S. S. \text{ due to (treat. + blocks)}$$

$$= S_3 \text{ (say)}$$

Now, we arrive at the following A. V. T.—

Source of variation	D. F.	S. S.	M. S. S.	F calculated	F tabulated at	
					5%	1%
Treatment	$v-1$ $=v_1$	S_1	$\frac{S_1}{v_1} = V_T$	$\frac{V_T}{V_E}$ if $V_T > V_E$
Block	$r-1$ $=v_2$	S_2	$\frac{S_2}{v_2} = V_B$	$\frac{V_B}{V_E}$ if $V_B > V_E$
Error	$(v-1)$ $(r-1)$ $=v_3$	S_3	$\frac{S_3}{v_3} = V_E$	—		
Totals	$vr-1$	S	—	—	—	—

If the calculated F corresponding to treatment comes out to be significant at $\alpha\%$ level, then we require farther analysis and compute the S. E. and C. D. as follows—

S. E. of the difference between two treatment means.

$$= \sqrt{\frac{2V_E}{r}}$$

$$(C. D.)_{\alpha\%} = (S. E. \text{ of difference}) \times t(\text{error d. f.})_{\alpha\%}$$

Inference:—The significant value of F leads to the conclusion that its corresponding factor has a significant effect on the yield and the treatment means differ significantly, they are arranged in the descending order of their magnitudes. Those pairs which do not differ significantly are underlined by a bar. Finally, the conclusions obtained from the analysis of the experimental data and comments on them are summarized in the form of a report.

Exp. (1):—Three varieties A, B and C were tested in a R. B. D. each with six replications. The layout and yields in lbs/plot are given in the diagram appended. Analyse the experimental yields and state your conclusions?—

I	II	III	IV	V	VI
A	C	B	C	A	B
17	35	21	25	33	19
C	B	C	A	B	A
33	23	29	17	23	23
B	A	A	B	C	C
19	29	25	15	37	27

Solution:—

Ho : Three varieties A, B and C and the blocks do not differ significantly in their yielding capacity.

For convenience of calculations, we take the deviations from $y=25\text{lbs}$ and arrange the data in the following table—

Block/ Treat.	I	II	III	IV	V	VI	Totals (T)	$(\text{Totals})^2$ = T^2	Mean
A	-8 (64)	4 (16)	0 (0)	-8 (64)	8 (64)	-2 (4)	-6 (212)	36	24
B	-6 (36)	-2 (4)	-4 (16)	-10 (100)	-2 (4)	-6 (36)	-30 (196)	900	20
C	8 (64)	10 (100)	4 (16)	0 (0)	12 (144)	2 (4)	36 (238)	1296	31
Totals (B)	-6 (164)	12 (120)	0 (34)	-18 (164)	18 (212)	-6 (44)	0=G (736)	0	
$(\text{Totals})^2$ = B^2	36	144	0	324	324	36	0		

$$C. F. = \frac{G^2}{N} = \frac{(0)^2}{18} = 0$$

$$T.S.S. = \sum \sum y_{ij}^2 - C. F. = 736 - 0 = 736$$

$$S. S. \text{ due to treat.} = \sum_i \frac{T_i^2}{r} - C. F. = \frac{36 + 900 + 1296}{6} - 0 = \frac{2232}{6} = 372.0$$

$$S. S. \text{ due to blocks} = \sum_j \frac{B_j^2}{v} - C. F. = \frac{36 + 144 + 0 + 324 + 324 + 36}{3} - 0 = \frac{864}{3} = 288.0$$

$$S. S. \text{ due to error} = T. S. S. - S. S. \text{ due to (treat. + blocks)} = 736 - (372 + 288) = 76.0$$

Now, we arrive at the following A. V. T.

Source of variation	D. F.	S. S.	M. S. S.	F call	F tab. at	
					5%	1%
Treat.	2	372.0	186.0	24.48**	4.10	7.56
Blocks	5	288.0	57.6	7.58**	3.33	5.64
Error	10	76.0	7.6	—	—	—
Totals	17	736.0	—	—	—	—

The value of F corresponding to treatment comes out to be highly significant, hence the treatment means differ significantly in their yielding capacity.

Now, in order to investigate which of the treatment pairs differ significantly, we compute the S. E. of the difference between the treatment-means and the C. D. as follows—

$$S. E. \text{ of the difference} = \sqrt{\frac{2V_E}{r}} = \sqrt{\frac{2 \times 7.6}{6}} = \sqrt{2.533} = 1.58$$

$$(C. D.) = (S. E. \text{ of difference}) \times t(10)_{5\%} = 1.58 \times 2.228 = 3.54 \text{ approx.}$$

'Now, we arrange the treatment means in the decreasing order of their yields—

Treat :	C	A	B
Mean :	31	24	20

Inference:—The three varieties *A*, *B* and *C* differ significantly at 1% level. The max. yield has been recorded in the case of variety *C* followed by variety *A*. The difference between *C* and *A* is significant at 5% level. *A* has also shown higher yield in comparison to *B* and the difference between them is also significant at 5%. Thus the variety *C* is the best of all as regards the average yield.

Exp. No. (2) : An experiment was carried out on wheat with three treatments in four randomized blocks. The plan and yield per plot in seers are given below—

Blocks			
I	II	III	IV
A 8	C 10	A 6	B 10
C 12	B 8	B 9	A 5
B 10	A 8	C 10	C 9

Analyse the data and state the conclusions ?

(M. Sc. Ag. Agra, 1959)

Solution—

Three treatments *A*, *B* and *C* and the blocks do not differ significantly.

For convenience of calculations, we take the deviations from $y=9$ seers, and arrange the data in the following tabular form—

Block treat.	I	II	III	IV	Totals =T	(Totals) ² =T ²	Mean
A	-1 (1)	-1 (1)	-3 (9)	-4 (16)	-9 (27)	81	6.75
B	1 (1)	-1 (1)	0 (0)	1 (1)	1 (3)	1	9.25
C	3 (9)	1 (1)	1 (1)	0 (0)	5 (11)	25	10.25
Totals = (B)	3 (11)	-1 (3)	-2 (10)	-3 (17)	-3=G (41)	9	
(Totals =B) ²	9	1	4	9	9		

$$C. F. = \frac{G^2}{N} = \frac{9}{12} = 0.75$$

$$T. S. S. = \sum_{i,j} y_{ij}^2 - C.F. = 41 - 0.75 = 40.25$$

$$S. S. \text{ due to treat.} = \sum_i \frac{T_i^2}{4} - C. F. = \frac{81+1+25}{4} - 0.75$$

$$= 26.75 - 0.75$$

$$= 26.00$$

$$S. S. \text{ due to blocks} = \sum_j \frac{T_j^2}{3} - C. F. = \frac{9+1+4+9}{3} - 0.75$$

$$= 7.67 - 0.75$$

$$= 6.92$$

$$S. S. \text{ due to error} = T. S. S. - S. S. \text{ due to (treat + blocks)}$$

$$= 40.25 - (26.00 + 6.92)$$

$$= 40.25 - 32.92$$

$$= 7.33$$

Now, we arrive at the following *A. V. T.*—

Sources of Variation	D. F.	S. S.	M. S. S	F Cal	F Tab at	
					5%	1%
Treatments	2	26.00	13.00	10.6*	5.14	10.92
Blocks	3	6.92	2.3067	1.8	4.76	9.78
Error	6	7.33	1.2222	—	—	—
Totals	11	40.25	—	—	—	—

The value of *F* corresponding to treatment comes out to be significant at 5% level and so the treatment means differ significantly in their yielding capacity.

Now, to decide which of the treatment pairs differ significantly, we compute the *S. E.* of the difference between two treatment means and *C. D.* as given below—

$$S. E. \text{ of the difference} = \sqrt{\frac{2V_E}{r}} = \sqrt{\frac{2 \times 1.22}{4}} = \sqrt{0.61} = 0.78$$

$$(C. D.) = (S. E. \text{ of the difference}) \times t(6)_{5\%} = 0.78 \times 2.447 = 1.91 \text{ approx.}$$

Now, we arrange the treatment means in the descending order of their magnitudes—

Treatment :	<i>C</i>	<i>B</i>	<i>A</i>
mean :	10.25	9.25	6.75

The treatment means which do not differ significantly from each other as regards their yields, are underlined by a bar.

Conclusion—The three treatments *A*, *B* and *C* differ significantly in their yielding-capacity. The max. yield is due to the treatment *C* followed by *B*, but their difference is not significant. The treatments *B* and *C* both differ significantly from the treatment *A*.

Exp. No. (3) : The per plot yields and a part of the analysis of variance of variety trial conducted in randomized blocks are given below—

Variety	Yields/plot				
	Blocks				
	I	II	III	IV	V
A	20	19	14	15	17
B	23	21	19	19	18
C	25	21	18	21	20
D	20	19	17	13	21

Analysis of Variance

Source of Variation	D. F.	S. S.
Blocks	4	72
Varieties	—	—
Error	—	—
Totals	19	158

Complete the analysis of variance ? What conclusions would you draw from the experiment ? (M. Sc. Ag. Agra, 1960)

Solution—

Ho : The four varieties A, B, C and D and the blocks do not differ significantly in their yielding capacity.

For convenience we take the deviations from $y=19$ and then arrange the data in the following tabular form—

Blocks Variety	I	II	III	IV	V	Totals T=	$(\text{Totals})^2$ (=T)	Mean
A	1 (1)	0 (0)	-5 (25)	4 (16)	-2 (4)	-10 (46)	100	17
B	4 (16)	2 (4)	0 (0)	0 (0)	-1 (1)	5 (21)	25	20
C	6 (36)	2 (4)	-1 (1)	2 (4)	1 (1)	10 (46)	100	21
D	1 (1)	0 (0)	-2 (4)	-6 (36)	2 (4)	-5 (45)	25	18

$$C. F. = \frac{G^2}{N} = \frac{(0)^2}{20} = 0$$

$$\sum \sum y^2 ij = (158) - 0$$

$$S. S. \text{ due to varieties} = \sum_i \frac{T_i^2}{5} - C. F. = \frac{100 + 25 + 100 + 25}{5} - 0$$

$$= \frac{250}{5} - 0 = 50$$

$$T. S. S. = \sum \sum y^2 ij - C. F. = 158 \text{ (given)}$$

$$S. S. \text{ due to error} = T. S. S. - S. S. \text{ due to (varieties + Blocks)}$$

$$= 158 - (50.0 + 12.0) = 158 - 122$$

$$= 36.0$$

Now, the analysis of variance table can be completed in the following manner—

Source of Variation	D. F.	S. S.	M S. S.	F cal.	F tab. at	
					5%	1%
Blocks	4	72.0	18.0	6.0**	3.26	5.1
Varieties	3	50.0	16.7	5.56*	3.49	5.95
Error	12	36.0	3.0	—	—	—
Totals	19	158.0	—	—	—	—

The calculated value of F corresponding to varieties comes out to be significant at 5% level while that corresponding to blocks comes as highly significant.

Thus, we conclude that the varieties A , B , C and D are significantly different in their yielding capacity at 5% level of significance.

In order to investigate which of the variety pairs differ significantly, we compute further the $S. E.$ of the difference between the two variety-means and the $C. D.$ as shown below—

$$S. E. \text{ of the difference} = \sqrt{\frac{2V_E}{r}} = \sqrt{\frac{2 \times 3}{5}} = \sqrt{1.2} = 1.094$$

$$(C. D.) = (S. E. \text{ of the difference}) \times t_{0.05} \quad (12)$$

$$\begin{aligned} 5\% &= 1.094 \times 2.179 = 2.3838 \\ &= 2.38 \end{aligned}$$

Now, we arrange the variety-means in their decreasing order of magnitudes —

Variety :	<i>C</i>	<i>B</i>	<i>D</i>	<i>A</i>
Mean :	21	20	18	17

The variety-means which do not differ significantly from each other at 5% level, are underlined by a bar.

Conclusion—The variety *C* produces the max. yield and followed by variety *B* but their difference is not significant at 5% level. The variety *C* differs significantly from *D* and *A* both. The variety *B* and *D* also *D* and *A* do not differ significantly and the variety *A* has the minimum (least) yielding-capacity. The varieties and the blocks differ significantly at 5% and 1% level of significance respectively.

Exp. No. (4) A varietal trial was conducted in a randomized block design with 9 varieties and 5 replications. In analysing the yield-data, the following sums of squares were obtained—

Blocks :	388.06
Varieties :	731.75
Error :	635.65

(a) Construct the analysis of variance and calculate the critical difference ?

(b) Arrange the varieties in order of performance. The variety means were—

Variety :	1	2	3	4	5	6	7	8	9
Yield in		21.40		18.50		24.90		14.00	
nds/acre :	20.80	21.40		19.80		15.40		11.30	

(M. Sc. Ag. Agra, 1963)

Solution—

Ho : The varieties and the blocks do not differ significantly.

(a) Under the above hypothesis, we prepare the following analysis of variance table—

Sources of Variation	D. F.	S. S.	M. S. S.	F. cal.	F. tab. at	
					5%	1%
Blocks	4	388.06	97.015	4.8**	2.674	3.982
Varieties	8	731.75	91.46875	4.6**	2.252	3.134
Error	32	635.65	19.88	—	—	—
Totals	44	1755.46	—	—	—	—

The calculated values of corresponding F_s indicate that the blocks as well as the varieties differ significantly in their yielding-capacity.

Now, the C. D. will be computed in the following manner—

$$S. E. \text{ of the difference} = \sqrt{\frac{2V_E}{r}} = \sqrt{\frac{2 \times 19.88}{5}} = \sqrt{7.952} = 2.82$$

$$(C. D.) = (S. E. \text{ of difference}) \times t(32)_{.05}$$

$$5\% = 2.82 \times 2.038 = 5.75$$

$$\text{and } (C. D.) = 2.82 \times t(32) = 2.82 \times 2.741$$

$$1\% \quad .01$$

$$= 7.73$$

(b) Variety : 6 2 3 1 5 4 7 8 9

Mean : 24.90 21.40 21.40 20.80 19.80 18.50 15.40 14.00 11.30

Inference—The above arrangement gives the idea that the variety no. six has the max. yielding capacity and the variety no. nine has the min. yielding capacity in the experiment performed.

EXERCISE II

Q. No. 1. Plan a varietal trial to test five improved varieties of wheat in order to select a suitable variety for your locality. Make use of R. B. D. with six replications. Construct the analysis of variance table and indicate how you would calculate the critical difference ?

(M. Sc. Ag. Agra, 1963)

Q. 2. (a) What considerations help you in determining the shape and arrangement of the blocks and plots in a field experiment using R. B. D. ?

(b) Give the relative merits and demerits of the R. B. D. over C. R. D. ?

(M. Sc. Ag. Agra, 1965)

Q. 3. (a) Randomize five treatments A, B, C, D and E to the plots of the following block ?

Plot	1	
	2	
	3	
	4	
	5	

(b) What should be the minimum number of replications to compare the two treatments v_1 & v_2 so that the d. f. for error be 12.

Ans. $r=13$

Q. 4. The yield in the plot for three varieties of maize each with 6 replications are given in the following table. Prepare the analysis of variance table and test the homogeneity between the three varieties A, B & C and their replications ?

Replications	VARIETIES		
	A	B	C
1.	252	213	199
2.	46	112	60
3.	29	133	165
4.	48	62	21
5.	10	27	154
6.	38	41	116

"

$$\text{Ans. } F = \frac{V_T}{V_E} = 1.49,$$

$$\text{and } F = \frac{V_R}{V_E} = 5.28$$

Q 5. (a) Define a R. B. D. and write down the procedure of randomization in a field experiment by taking a suitable example ?

(b) A district N.C.C. head quarter appointed five N.C.C. officers A, B, C, D & E each to five degree colleges in the district to impart a certain Military-training to the N.C.C. cadets in the institutions. The no. of cadets as trainees under the guidance of each officer in the five colleges are recorded below. Test the homogeneity of the data with respect to the N.C.C. officers and the different colleges ?

Officers	Colleges				
	I	II	III	IV	V
A	50	70	115	95	100
B	55	75	120	85	105
C	45	60	100	80	95
D	40	65	85	85	100
E	60	75	95	75	110

Ans. (b) $F=47.0$ for colleges
and $F=2.99$ for officers

Q. 6. A farmer grouped into 4, his 24 cows of 6 breeds and fed them with 4 rations A, B, C & D for a fortnight. Then the increase in milk-yield in ounces/cow, were recorded as given in the following table. Analyse the data and test, whether there is any significant difference between the four rations and breeds at 5% level of significance ? Given that $F_{.05}(3, 15)$ & $F_{.05}(5, 15)$ are 3.29 & 5.05 respectively.

Rations	Breeds					
	I	II	III	IV	V	VI
A	20	22	20	22	24	24
B	26	24	20	24	22	22
C	24	22	22	26	24	22
D	26	28	24	30	26	22

Ans. $F=2.23$ for breeds
and $F=5.03$ for rations

Q. 7. Below are given the marks obtained by 12 candidates; appeared in a P.S.C. interview held on 4 current topics by a committee of 3 experts, each handling four candidates for the different current topics. Analyse the data and test for the homogeneity between the interviewers and the four current topics ?

Interviewers	Topics			
	I	II	III	IV
A	8.0	26.0	15.0	21.0
B	22.0	25.0	24.0	36.0
C	11.0	27.0	13.0	21.0

Ans. $F=5.76$ for interviewers
and $F=6.2$ for topics

Q. 8. Following table gives the number of pods/plant for 20 pea-plants of 4 different varieties v_1, v_2, v_3, v_4 . Test the significance of the difference between the varieties and the plants ?

Variety	Plants				
	I	II	III	IV	V
V_1	23	26	30	22	25
V_2	29	23	27	30	26
V_3	21	23	26	24	25
V_4	25	22	29	27	28

Ans. $F=2.03$ for plants
and $F=1.67$ for varieties

CHAPTER IV

Latin Square Design (L. S. D.)

Description:—In R.B.D , the whole experimental area is divided into homogeneous blocks and randomization is kept restricted within the blocks *i.e.* subject to one restriction only. While in L.S.D. the experimental-area is divided into rows and columns such that the no. of rows and columns is equal and each treatment (denoted by Latin letters) occurs only once in a row and column. Thus the randomization is subject to 2 restrictions. This arrangement removes (eliminates) the variation between the rows and the columns from that of within and reduces the experimental-error considerably.

Applications :—In the field experimentation, it is used :—

(i) When the fertility-gradient is in one direction but not known.

(ii) When the fertility gradient is in two directions at right-angles.

In fact, the L.S.D. can be applied in all the cases where the variation in the experimental-material is from two orthogonal sources. It is used in industry, animal-husbandary, piggery, green house, biological and social sciences where it is desired to control simultaneously two factors contributing to the experimental-error.

Relative merits and Demerits of L.S.D. over a R.B.D. :—
Although the L. S. D. is an improvement over the R.B.D. but there are situations where R. B. D. is used instead of L.S.D. :—

(i) R.B.D. can be used with any no. of treatments and their replications, but L.S.D. is a suitable design for the no. of treatments from 5 to 12.

(ii) There is no restriction on the no. of replications in a R.B.D. while the no. of treatments and replications should be equal in a L S D. This restriction puts a limit on its applications.

(iii) In the case of missing-plots, the statistical-analysis is simple in a R.B.D. but in L.S.D. it becomes some what complex and

especially when the missing units are several.

(iv) In the field, the R.B.D. is easier to manage than a L.S.D. As it can be performed equally in a rectangular or square field or a field of any other shape, while the L.S.D. necessitates approximately a square field.

The merits of the L. S. D. lie in the fact that it controls simultaneously two factors contributing to the experimental error. Thus the L. S. D. is a more suitable design than a R. B. D. where the fertility gradient is in to 2 directions at right angles or in one unknown direction.

Replications:—In a L. S. D. the no. of replications must be equal to the no. of treatments. Due to this fact, the design is not suitable for a large no. of treatments and is rarely used for the no. of treatments greater than 12. On the ground of error d. f., it is not suitable for fewer treatments also i. e. < 5 . The error—d. f. for a 2×2 , 3×3 and 4×4 L. S. D. are 0, 2 and 6 respectively. According to Prof. R. A. Fisher, it is most suitable for the no. of treatments from 5 to 8 and can be used up to 12.

Randomization:—(a) In the randomization of a $K \times K$, L. S. D. ($K \leq 4$), the first step is to select a reduced L. S. (Standard L. S.) from the set of reduced L. squares. A standard L. S. is one which has an alpha-betical order of letters in the first row and first column.

For the 2×2 , 3×3 , and 4×4 , L. square, the standard (reduced) L. Squares are—1, 1, 4 and they are—

For 2×2

A	B
B	A

For 3×3

A	B	C
B	C	A
C	A	B

For 4×4

(i)

A	B	C	D
B	C	D	A
C	D	A	B
D	A	B	C

(ii)

A	B	C	D
B	A	D	C
C	D	B	A
D	C	A	B

(iii)

A	B	C	D
B	A	D	C
C	D	A	B
D	C	B	A

(iv)

A	B	C	D
B	D	A	C
C	A	D	B
D	C	B	A

After the selection of a reduced L. S., the second step is to randomize the order of all the 'K' columns and the last (K-1) rows.

(b) For the randomization of a $K \times K$ ($K \geq 5$) L. S., construct any L. S. in the first step and then randomize the order of its all rows, columns and treatments (letters). The detailed procedure for a 5×5 L. S. is given below—

(i) Construct a 5×5 L. S., let it be.

Row No.

1	A	B	C	D	E
2	B	A	E	C	D
3	C	D	A	E	B
4	D	E	B	A	C
5	E	C	D	B	A

(ii) Randomize the order of the above L. S. Let the random numbers be 4, 3, 2, 5 and 1. According to these numbers, the fourth row of the above L. S. will be written in place of first row, third in place of second, second in place of third, 5th in place of 4th and first in place of 5th. Arranging them in this order, we have.

Col. No

1	2	3	4	5
D	E	B	A	C
C	D	A	E	D
B	A	E	C	D
E	C	D	B	A
A	B	C	D	E

(iii) Randomize the order of the columns in the order of the random numbers 3, 1, 4, 5, 2 (a fresh set of random numbers). Then we obtain—

B	D	A	C	E
A	C	E	B	D
E	B	C	D	A
D	E	B	A	C
C	A	D	E	B

(iv) Finally, randomize the order of Latin-letters. Let the set of a freshly selected random numbers be 2, 1, 5, 3 and 4. According to this set, B will be written for A, A for B, E for C, C for D and D for E. Reshuffling the letters in this order, we get---

A	C	B	E	D
B	E	D	A	C
D	A	E	C	B
C	D	A	B	E
E	B	C	D	A

Layout:—In field experiments, L. S. is generally performed on a square or nearly square area but it is not necessary. As the object of this design is to control variation in two directions, it can be applied when the plots are in a row or set of rows. In this case, the compact blocks are supposed to be the rows of a L. S. and the order of plots within the blocks the columns. This type of arrangement is usually done in a green house experimentation.

Formulation of Null Hypothesis:—In order to compare the

treatments in a L. S. D., we set up the null hypothesis *-that the data is homogeneous i. e, the variation is the data in due to chance (error) only.*

Statistical Analysis:—To reduce the bulk of calculations, generally we take the deviations of the yields from some appropriate origin in a ' $K \times K$ ' L. S. D. and then obtain the totals for rows, columns and treatments, which are usually denoted by (R_1, R_2, \dots, R_K) , (C_1, C_2, \dots, C_K) and (T_1, T_2, \dots, T_K) respectively. Then the sum of squares (S. S.) for rows, columns and treatments are computed by the following formulæ—

$$T. S. S. = \sum_i \sum_j y_{ij}^2 - C. F., \text{ where } i = 1, 2, \dots, k \text{ and } C. F. = \frac{G^2}{K^2}$$

$$= S \text{ (say)} \quad (K^2 - N)$$

y_{ij} → stands for the yield of the plot in i th row and j th. Column for specified treatment written above it by Lattin-letter.

$$S. S. \text{ due to rows} = \sum_i \frac{R_i^2}{K} - C. F.,$$

$$= S_1 \text{ (say)}$$

Where R_i → stands for the total of i th row.

$$S. S. \text{ due to columns} = \sum_j \frac{C_j^2}{K} - C. F.,$$

$$= S_2 \text{ (say)}$$

Where C_j → stands for the total of j th column.

$$S. S. \text{ due to treatments} = \sum_i \frac{T_i^2}{K} - C. F.,$$

$$= S_3 \text{ (say)}$$

Where T_i → stand for the total of i th. ($i = 1, 2, \dots, k$) treatment.

$$S. S. \text{ due to error} = T. S. S. - S. S. \text{ due to (rows + columns + treatment)}$$

$$= S - (S_1 + S_2 + S_3)$$

$$= S_4 \text{ (say)}$$

Now we arrive at the following analysis of variance table :—

Source of variation	D.F.	S.S.	M.S.S.	F. cal.	F tab. at	
					5%	1%
Rows	$K-1=v_1$	S_1	$\frac{S_1}{v_1} = V_R$	$\frac{V_R}{V_E}$ if $V_R > V_E$
Columns	$K-1=v_1$	S_2	$\frac{S_2}{v_1} = V_C$	$\frac{V_C}{V_E}$ if $V_C > V_E$
Treatments	$K-1=v_1$	S_3	$\frac{S_3}{v_1} = V_T$	$\frac{V_T}{V_E}$ if $V_T > V_E$
Error	$(K-1)(K-2)$ $=v_2$	S_4	$\frac{S_4}{v_2} = V_E$
Totals	K^2-1	S

If the value of F corresponding to the treatments comes out to be significant at 5% or at 1% level of significance, we require a further analysis of the treatments to decide the significance of the difference between the two treatments. For this purpose, we compute the S.E. of the difference of the treatment-means and the C.D. at the same level, as given below :—

S. E. of the difference between two treatment means

$$= \sqrt{\frac{2V_E}{K}} \quad \text{and}$$

$$\begin{aligned} (C. D.) &= (S. E. \text{ of the difference}) \times t(\text{error d. f.}) \\ \alpha\% &\qquad \qquad \qquad \alpha\% \end{aligned}$$

Inference :—Any significant value of F leads to the conclusion that the corresponding factor has a significant effect on the plot-yield. If the treatments differ significantly then their means are arranged in decreasing order of their magnitudes. The treatment mean-pairs, for which the differences are less than the C.D. are underlined by a bar indicating that their differences are not significant or the two do not differ significantly from each other at the given level of significance.

Finally, the conclusions and the comments on them are summarized in the form of a report.

Exp. (1) Carryout the analysis of the following L.S.D. :—

50	70	70	80	90
A	B	C	D	E
70	90	80	80	50
B	C	D	E	A
60	50	90	80	90
C	D	E	A	B
50	60	80	50	70
D	E	A	B	C
80	90	50	70	60
E	A	B	C	D

(B. A. Vikram, 1961)

Solution :—

Ho : The data is homogeneous.

For the convenience of calculations, we shift the origin to $y=70$ and prepare the following table to compute the sum of squares :—

Row \ Col. ⁿ	1	2	3	4	5	Total=R	(Total=R) ²
1	A -20 (400)	B 0 (0)	C 0 (0)	D 10 (100)	E 20 (400)	10 (900)	100
2	B 0 (0)	C 20 (400)	D 10 (100)	E -20 (400)	A -20 (400)	20 (1000)	400
3	C -10 (100)	D -20 (400)	E 20 (400)	A 10 (100)	B 20 (400)	20 (1400)	400
4	D -20 (400)	E -10 (100)	A 10 (100)	B -20 (400)	C 0 (0)	-40 (1000)	1600
5	E 10 (100)	A 20 (400)	B -20 (400)	C 0 (0)	D -10 (100)	0 (1000)	0
Total=C	-40 (1000)	10 (1300)	20 (1000)	10 (700)	10 (1300)	10=G (5300)	100
(Total=C) ²	1600	100	400	100	100	100	
Treat.	A 0 (0)	B -20 (100)	C 10 (100)	D -30 (900)	E 50 (2500)		
Mean	70	66	72	64	80		

$$C.F. = \frac{G^2}{n} = \frac{100}{25} = 4.0$$

$$T.S.S. = \sum \sum y_{ij}^2 - C.F. = 5300 - 4.0 = 5296.0$$

$$S.S. \text{ due to rows} = \sum_i \frac{R_i^2}{5} - C.F. = \frac{100 + 400 + 400 + 1600 + 0}{5} - 4.0$$

$$= \frac{2500}{5} - 4.0 = 500 - 4.0 = 496.0$$

$$S.S. \text{ due to cols.} = \sum_j \frac{C_j^2}{5} - C.F. = \frac{1600 + 100 + 400 + 100 + 100}{5} - 4.0$$

$$= \frac{2300}{5} - 4.0 = 460 - 4.0 = 456.0$$

$$S.S. \text{ due to treat.} = \sum_i \frac{T_i^2}{5} - C.F. = \frac{0 + 400 + 100 + 900 + 2500}{5} - 4.0$$

$$= \frac{3900}{5} - 4.0 = 780 - 4.0 = 776.0$$

$$S.S. \text{ due to error} = T.S.S. - S.S. \text{ due to (rows + cols. + treat.)}$$

$$= 5296 - (496 + 456 + 776) = 5296 - 1728$$

$$= 3568$$

Now we arrive at the following A.V.T. :-

Source of variation	D. F.	S. S	M. S. S	F. cal.	F tab. at	
					5%	1%
Rows	4	496	124	$\frac{297.33}{124} = 2.39$	5.91	14.37
Columns	4	456	114	$\frac{297.33}{114} = 2.60$	"	"
Treat.	4	776	194	$\frac{297.33}{194} = 1.53$	"	"
Error	12	3568	297.33	—	—	—
Totals	24	5296	—	—	—	—

The calculated values of F corresponding to rows, columns and treatments indicate that none of them is significant at 5% level of significance. Thus the L.S.D. provides no improvement over a C.R.D. in this case.

Inference :—The data is homogeneous.

Exp. 2. A varietal trial was conducted on wheat with four varieties in a L.S.D. The plan of the experiment and the per plot yield are given below :—

C 25	B 23	A 20	D 20
A 19	D 19	C 21	B 18
B 19	A 14	D 17	C 20
D 17	C 20	B 21	A 15

Analyse the data and interpret the results ?

(M. Sc. Ag. Agra, 1961)

Solution :—

The data is homogeneous.

To carryout the analysis, we take the deviations from $y=20$ for our convenience in calculations and arrange the data in the following tabular form :—

Row. \ Col. ⁿ	1	2	3	4	Totals R	(Totals = R) ²
1	C 5 (25)	B 3 (9)	A 0 (0)	D 0 (0)	8 (34)	64
2	A -1 (1)	D -1 (1)	C 1 (1)	B -2 (4)	-3 (7)	9
3	B -1 (1)	A -6 (36)	D -3 (9)	C 0 (0)	-10 (46)	100
4	D -3 (9)	C 0 (0)	B 1 (1)	A -5 (25)	-7 (35)	49
Totals = C	0 (36)	-4 (46)	-1 (11)	-7 (29)	-12 (122)	144
(Totals = C) ²	0	16	1	49	144	
Variety	A -12 (144)	B 1 (1)	C 6 (36)	D -7 (49)		
Mean	17	20.25	21.5	18.25		

$$C. F. = \frac{G^2}{N} = \frac{144}{16} = 9.0$$

$$T.S.S. = \sum \sum y_{ij}^2 - C.F. = 122 - 9.0 = 113.0$$

$$S.S. \text{ due to rows} = \sum \frac{R_i^2}{4} - C.F. = \frac{64+9+100+49}{4} - 9.0$$

$$= \frac{222}{4} - 9.0 = 55.5 - 9.0 = 46.5$$

$$S.S. \text{ due to col}^n s. = \sum \frac{C_j^2}{4} - C.F. = \frac{0+16+1+49}{4} - 9.0$$

$$= \frac{66}{4} - 9.0 = 16.5 - 9.0 = 7.5$$

$$S.S. \text{ due to varieties} = \frac{T_A^2 + T_B^2 + T_C^2 + T_D^2}{4} - C.F.$$

$$= \frac{144+1+36+49}{4} - 9.0$$

$$= \frac{230}{4} - 9.0 = 57.5 - 9.0 = 48.5$$

$$S.S. \text{ due to error} = T.S.S. - S.S. \text{ due to (rows + Col}^n s. + \text{varieties)}$$

$$= 113 - (46.5 + 7.5 + 48.5) = 113 - 102.5$$

$$= 10.5$$

Now we arrive at the following A.V.T. :—

Source of variation	D. F	S. S.	M. S. S.	F. cal.	F .05	F .01
Varieties	3	48.5	16.167	9.23*	4.76	9.78
rows	3	46.5	15.5	8.85*	„	„
columns	3	7.5	2.5	1.42	„	„
Error	6	10.5	1.75	—	—	—
Totals	15	113.0	—	—	—	—

The calculated values of F_s corresponding to varieties and rows indicate that the varieties and the rows are significantly different at 5% level of significance. The insignificant value of F corresponding

to columns indicates that the L.S.D. is not an improvement over R.B.D. in this case.

Now to determine which of the variety-pairs differ significantly, we require the computations of S.E. of the difference of means of varieties and the C.D.

$$\begin{aligned} \text{S.E. of mean difference} &= \sqrt{\frac{2\bar{V}_E}{r}} = \sqrt{\frac{2 \times 1.75}{4}} \sqrt{0.875} \\ &= 0.935 \end{aligned}$$

$$\begin{aligned} (C.D.)_{5\%} &= (\text{S.E. of difference}) \times t_{(6)} \\ &= 0.935 \times 2.447 = 2.879 \\ &\approx 2.29 \end{aligned}$$

Now, we arrange the variety-means in their decreasing order of magnitudes :

Variety :	C	B	D	A	
Mean :	21.50	20.25	18.25	17.00	The varieties

Which do not differ significantly from each other, are under-lined by a bar.

Conclusion :—The varieties have significant effect on yield at 5% level and the variety C has max. yielding capacity followed by variety B but it does not differ significantly from B. The variety C differs significantly from each of D and A. The variety B also does not differ significantly from D but differs significantly from A. We also note that the varieties D and A do not differ significantly as regards their average yield.

Exercise III

Q. 1. (a) Compare the advantages and disadvantages of the R.B.D. and L.S.D. in field trials. (M. Sc. Ag. Agra, 1960, 62)

(b) In a trial of five varieties of wheat, A, B, C, D and E laid out in a Latin square, the following yields (in oz./plot.) were obtained :—

B	E	C	A	D
36	56	164	120	80
E	D	B	C	A
66	64	76	178	60
C	A	D	B	E
118	76	70	64	34
A	C	E	D	B
58	146	66	48	40
D	B	A	E	C
60	16	66	66	72

Analyse the data and state your conclusions ?

(Hint : take deviations from 76)

$$\text{Ans. (b) } F = \frac{V_T}{V_E} = 27.10, \quad F = \frac{V_R}{V_E} = 4.44 \text{ and}$$

$$F = \frac{V_C}{V_E} = 5.18$$

Q. 2. Five varieties A, B, C, D and E of millet were tested in a 5×5 L.S.D. The layout and the yield in lbs/plot are given below. Analyse the data and test for the variation between the different varieties ?

A	B	E	C	D
60	18	28	82	40
B	A	C	D	E
38	30	89	32	33
E	C	D	B	A
17	59	35	32	38
C	D	A	E	B
73	24	29	33	20
D	E	B	A	C
30	33	8	33	36

(Hint : take deviations from 38)

$$\text{Ans. } V_T = 1531.7, \quad V_C = 123.7, \quad V_R = 258.5 \text{ and}$$

$$V_E = 149.2$$

Q. 3 (a) What is a L.S.D. ? Give the assumptions and applications of a L.S.D. in field experimentations ?

(b) A 4×4 L.S. was laid out to test the effects of various fertilizers on the yield of potatoes. Here is the field-plan with the plot yield (bushels/acre).

The letters specify the treatments. Analyse the data and draw your conclusions ?

Col. Row	1	2	3	4
1	A 423	B 428	C 452	D 390
2	B 425	D 380	A 420	C 440
3	C 460	A 414	D 375	B 425
4	D 380	C 450	B 430	A 412

(Hint : Shift the origin to 419)

Ans. $V_T = 3312.5$, $V_C = 20.167$, $V_R = 35.833$ and

$$V_E = 41.75$$

Q. 4. At a biological research centre, a research assistant fed up the rabbits of four different breeds for a month with 4 types of rations (A, B, C & D) and noted the gains in their weights in ounces. He presented the results of weights in a 4×4 L.S. The layout and the gains in weights are given in the following table :—

A 5.0	B 12.0	C 13.1	D 8.0
B 13.5	D 10.5	A 14.0	C 12.0
C 14.0	A 8.5	D 10.0	B 13.0
D 7.5	C 15.5	B 11.1	A 11.5

Analyse the data and write a brief report ?

(Hint : take the deviations from 11.2)

Ans. $V_R = 6.09$, $V_C = 3.13$, $V_T = 19.18$ and $V_E = 5.68$

Q. 5. The following data shows the results of a varietal trial on wheat in a Latin Square. The varieties are designated A, B, C, D, E and F. The yield of different plots are written below the treatments.

Analyse the data and give your inferences ?

Plan and yields of a varietal trial on wheat in a L. S. D.

E 433	B 327	F 452	A 190	C 304	D 216
B 289	C 275	D 215	E 288	F 371	A 82
A 184	E 281	C 283	B 222	D 134	F 446
F 420	D 248	E 305	C 239	A 123	B 184
D 252	A 232	B 211	F 417	E 394	C 266
C 300	F 305	A 59	D 166	B 126	E 220

Ans. $V_R = 10839.71$, $V_C = 4893.45$

$V_T = 49635.984$ and $V_E = 1527.0515$

Q. 6. Carry out the analysis of variance of the following data in a 3×3 Latin Square ?

Col: Row	1	2	3
1	A 122	B 148	C 132
2	C 131	A 127	O 139
3	B 128	C 129	A 119

Ans. $F = \frac{V_T}{V_E} = 10.7$

Q. 7. Describe an experiment for comparing the effect of 5 different feeds on milk-yield of dairy cows using a Latin Square Design. Give full details of plan and conduct of the experiment and explain the method of analysis of results ?

(M. Sc. Ag. Agra, 1956)

Hint :—Let the 5 diets to be compared be A, B, C, D and E. For conducting the experiment in a 5×5 L. S., we require in all 25 dairy cows as our experimental material. To satisfy the conditions of L. S., these 25 cows should be different with respect to two factors of variation e.g. the breed and age-group. Let the 5 breeds be b_1, b_2, b_3, b_4 and b_5 and the five age-groups be 3-5, 5-7, 7-9, 9-11 & 11 and over years. There are 5 cows in each age-group and of each breed such that no two cows of the same breed are of the same age-group. Each of the five diets will be given to 5 cows such that no two cows receiving the same diet are of the same breed and of the same age-group. The classification of cows with respect to age-group is represented here by columns and breeds by rows and the diets A, B, C, D, and E are assigned in such a way that each occurs once in a row and a column. Thus the arrangement will be as follows :—

Col. Row	1 3-5	2 5-7	3 7-9	4 9-11	5 11 & over
b_1	A 1	B 2	E 3	C 4	D 5
b_2	B 6	A 7	C 8	D 9	E 10
b_3	E 11	C 12	D 13	B 14	A 15
b_4	C 16	D 17	A 18	A 19	B 20
b_5	D 21	E 22	B 23	A 24	C 25

For randomization and Statistical analysis see theory.

The figures written below the letters denote the cow-number.

CHAPTER V

Analysis of Covariance

In the previous chapters, we have seen how the experimental error is reduced by increasing the no. of replications, refining the experimental technique and grouping the experimental-material into homogenous groups (blocks). There is yet one more method which is found very useful in reducing the experimental-error. This method eliminates the contribution made by the uncontrolled factors (related to the yield) to the experimental error. This elimination is possible only, when the uncontrolled factors can be measured quantitatively. Thus the error can be controlled by measuring such characters (factors) in addition to the yields, the characteristic statistical-tool for this control is the *Covariance analysis*. The Covariance-analysis makes the use of the regression of yield on the related uncontrolled-factor for making a correction in the estimates of treatment-differences and experimental error. This control of error by analysis of covariance-technique is called the *Statistical control of error*. According to Prof. 'R. A. Fisher', "*analysis of covariance combines the advantages and reconciles the requirements of the two very widely and applicable procedures known as regression and analysis of variance.*"

In the field experimentation, the no. of plants, tillers, prunings, and age of the crop, etc. are the extraneous sources of variation which contribute to the experimental error. The effect of such a character vary from plot to plot randomly. The variate 'x' associated with the observation of the uncontrolled factor is called the 'Concomitant or ancillary variate.' This variate should be such that it remains unaffected by the treatments, other wise the results will be misleading.

Applications:—

(i) One of the most important applications of the analysis of covariance is the control of errors that arise at random, known as '*statistical control of error*'.

(ii) Covariance-analysis can be applied in sorting out the regression effect.

(iii) The analysis of covariance also provides a unique method to test the significance of the difference between the two or more regression coefficients.

(iv) Covariance-analysis can successfully be used to carry out the analysis of the incomplete data (when one or more units are missing in the data).

(v) The technique of covariance-analysis is more effective in reducing the experimental error than the process of grouping the experimental material into homogeneous groups as the former provides more error d. f. than the later.

Assumptions:—The assumptions involved in the analysis of covariance are—

(i) The effects of different factors i. e. treatments, groups and regression are additive.

(ii) Apart from the regression-effect, the yields are distributed normally and independently.

(iii) The concomitant variate 'x' is not affected by the treatment.

Now, we give the procedure of statistical control of error using a single concomitant variate in the case of C. R. D., R. B. D, and L. S. D.

(1) Case of C.R.D.

Suppose, we have got ' v ' treatments, the i th treatment being replicated ' r_i ' times ($i=1, 2, \dots, v$) and the variate ' y ' denotes the yield while ' x ' the concomitant variate. The data from the original layout can be arranged in the following tabular form—

Replication \ Treatments	1		2		v			
	y	x	y	x		y	x		
1	y_{11}	x_{11}	y_{21}	x_{21}	y_{v1}	x_{v1}		
2	y_{12}	x_{12}	y_{22}	x_{22}	y_{v2}	x_{v2}		
\vdots	\vdots	\vdots	\vdots	\vdots		\vdots	\vdots		
\vdots	\vdots	\vdots	\vdots	\vdots		\vdots	\vdots		
r_i	y_{1r_1}	x_{1r_1}	y_{2r_2}	x_{2r_2}		y_{vr_v}	x_{vr_v}		
Totals	$T_{1(y)}$	$T_{1(x)}$	$T_{2(y)}$	$T_{2(x)}$	$T_{v(y)}$	$T_{v(x)}$	$G_{(y)}$	$G_{(x)}$

Total no. of pairs = $N = \sum r_i$

where $T_{i(y)}$ and $T_{i(x)}$ stand for i th treatment totals for ' y ' and ' x ' variates respectively. Further steps are as follows—

(i) Compute the $S. S._x$ for treatments and error by the following formulae—

$$[a] \quad T. S. S._x = \sum_i \sum_j x_{ij}^2 - C. F._x \quad \text{where } C. F._x = \frac{G_{(x)}^2}{N}$$

$$[b] \quad S. S._x \text{ (treat.)} = \sum_i \frac{T_{i(x)}^2}{r_i} - C. F._x, \\ = A_1 \text{ (say)}$$

$$\text{and } [c] \quad S. S._x \text{ (error)} = T. S. S._x - S. S._x \text{ (treat.)} \\ = A \text{ (say)}$$

ii) Compute the $S. S._y$ for treatment and error by the following formulae—

$$[a] \quad T. S. S._y = \sum_i \sum_j y_{ij}^2 - C. F._y, \quad \text{where } C. F._y = \frac{G_{(y)}^2}{N}$$

$$[b] \quad S. S._y \text{ (treat.)} = \sum_i \frac{T_{i(y)}^2}{r_i} - C. F._y, \\ = B_1 \text{ (say)}$$

$$\text{and } [c] \quad S. S._y \text{ (error)} = T. S. S._y - S. S._y \text{ (treat.)} \\ = B \text{ (say)}$$

(iii) Compute the sum of products $S. P_{xy}$ for treatment and error by the following formulae--

$$[a] T. S. P_{xy} = \sum_i \sum_j x_{ij} y_{ij} - C. F_{xy},$$

$$\text{where } C. F_{xy} = \frac{G_{(x)} G_{(y)}}{N}$$

$$[b] S. P_{xy} (\text{treat.}) = \sum_i \frac{T_i(x) T_i(y)}{r_i} - C. F_{xy},$$

$$= C_1 \text{ (say)}$$

$$\text{and } [c] S. P_{xy} (\text{error}) = T. S. P_{xy} - S. P_{xy} (\text{treat.})$$

$$= C \text{ (say)}$$

(iv) Summarize the above results in the following tabular form--

Source of variation	$D. F.$	$S. S_x$	$S. S_y$	$S. P_{xy}$	Reg. coefft. 'b'
Treatment	$v-1=v_1$	A_1	B_1	C_1	—
Error	$N-v=v$	A	B	C	C/A
Totals $= (T+E)$	$N-1$ $= v_1+v$	(A_1+A)	(B_1+B)	(C_1+C)	—

(v) Test the significance of the regression coefficient 'b' = C/A in the error line at 5% by computing the statistic--

$$F(1, v-1) = \frac{C^2/A}{B - C^2/A} \times (v-1) \quad \text{If this cal. } F \text{ comes out to}$$

be significant, the concomitant variate has an effect on the yield and covariance-analysis will be used to eliminate its effect. On the other hand, if F is not significant, the consideration of the concomitant variate is useless and the analysis of variance will be used to compare the treatments.

(vi) In case when F is significant, obtain the adjusted $S. S_y$ in the following way--

$$[a] \text{ adjusted } T. S. S_y = (B_1+B) - \frac{(C_1+C)^2}{(A_1+A)} = S \text{ (say)}$$

$$[b] \text{ adjusted } S. S_y (\text{error}) = B - C^2/A = S_2 \text{ (say)}$$

$$[c] \text{ adjusted } S. S._v (\text{treat.}) = \text{adjusted } T. S. S._v - \text{adjusted } S. S._v (\text{error})$$

$$= S. S._v (\text{say})$$

(viii) Prepare the following A. V. T.—

Source of variation	<i>D. F.</i>	<i>S. S.</i> adj	<i>M. S. S.</i>	<i>F</i> cal.	<i>F</i> tab. at.	
					5%	1%
Treatment	ν_1	S_1	$\frac{S_1}{\nu_1} = V' T$	$\frac{V' T}{V' E}$ if $V' T > V' E$
Error	$\nu - 1$	S_2	$\frac{S_2}{\nu - 1} = V' E$	—	—	—
Totals	$N - 2$	S	—	—	—	—

$$S. E. \text{ of the difference} = \sqrt{V' E \left[\frac{1}{r_i} + \frac{(\bar{y}_i - \bar{y}_j)^2}{A} + \frac{1}{r_j} \right]},$$
$$(C. D.) = \frac{(S. E. \text{ of difference}) \times t}{.05} \quad (v-1)$$

(x) Adjust the mean-yields by the following formulae—

$$y'_i = y_i - b(\bar{x}_i - \bar{x}), \text{ where } b = C/A, \bar{x} = \frac{G(x)}{N}$$

\bar{y}'_i is the adjusted mean yield for i th treatment and $\bar{y}_i = \frac{T_i(y)}{r_i}$,

$$\bar{x}_i = \frac{T_{i(x)}}{T_i}$$

Arrange the adjusted mean-yields in the descending order of their magnitudes and under line those pairs by a bar which do not differ significantly from each other.

(xi) Finally summarize the results obtained and comment on them if any, in the form of a report.

(2) Case of R. B. D.

Suppose we have got ' v ' treatments each replicated ' r ' times. The variate ' y ' denotes the yield and ' x ' the concomitant variate. The data from the original layout can be arranged in the following tabular form—

Blocks \ Treat	1				v		Totals	
	y	x	y	x		y	x	$B_{(y)}$	$B_{(x)}$
1	y_{11}	x_{11}	y_{21}	x_{21}	y_{v1}	x_{v1}	$B_{1(y)}$	$B_{1(x)}$
2	y_{12}	x_{12}	y_{22}	x_{22}	y_{v2}	x_{v2}	$B_{2(y)}$	$B_{2(x)}$
⋮	⋮	⋮	⋮	⋮		⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮		⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮		⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮		⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮		⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮		⋮	⋮	⋮	⋮
r	y_{1r}	x_{1r}	y_{2r}	x_{2r}	y_{vr}	x_{vr}	$B_{r(y)}$	$B_{r(x)}$
Totals	$T_{1(y)}$	$T_{1(x)}$	$T_{2(y)}$	$T_{2(x)}$	$T_{v(y)}$	$T_{v(x)}$	$G_{(y)}$	$G_{(x)}$

where $B_{j(y)}$ and $B_{j(x)}$ stand for the j th block totals for the variates ' y ' and ' x ' respectively.

Further steps are as follows—

(i) Compute the $S. S_x$ for treatments, blocks and error as given below—

$$[a] T. S. S_x = \sum_i \sum_j x_{ij}^2 - C. F_x \text{ where } C. F_x = \frac{G^2(x)}{N} \quad \text{and } N = vr$$

$$[b] S. S_x (\text{treat.}) = \sum_i \frac{T_{i(x)}^2}{r} - C. F_x = A_1 \text{ (say) ,}$$

$$[c] S. S_x (\text{blocks}) = \sum_j \frac{B_{j(x)}^2}{v} - C. F_x = A_2 \text{ (say) ,}$$

$$\text{and } [d] S. S_x (\text{error}) = T. S. S_x - S. S_x (\text{treat.} + \text{blocks}) \\ = T. S. S_x - (A_1 + A_2) = A \text{ (say)}$$

(ii) Compute the $S. S_y$ for treatments, blocks and error by the formulae—

$$[a] T. S. S_y = \sum_i \sum_j y_{ij}^2 - C. F_y, \quad \text{where } C. F_y = \frac{G^2(y)}{N} ,$$

$$[b] S. S_y (\text{treat.}) = \sum_i \frac{T_{i(y)}^2}{r} - C. F_y = B_1 \text{ (say) ,}$$

$$[c] S. S_y (\text{blocks}) = \sum_j \frac{B_{j(y)}^2}{v} - C. F_y = B_2 \text{ (say) ,}$$

$$\text{and } [d] S. S_y (\text{error}) = T. S. S_y - S. S_y (\text{treat.} + \text{blocks}) \\ = T. S. S_y - (B_1 + B_2) = B \text{ (say)}$$

(iii) Compute $S. P_{xy}$ for treatments, blocks and error by the following formulae—

$$[a] T. S. P_{xy} = \sum_i \sum_j x_{ij} y_{ij} - C. F_{xy}, \quad \text{where } C. F_{xy} = \frac{G(x)G(y)}{N} ,$$

$$[b] S. P_{xy} (\text{treat.}) = \sum_i \frac{T_{i(x)} T_{i(y)}}{r} - C. F_{xy} = C_1 \text{ (say) ,}$$

$$[c] S. P_{xy} (\text{blocks}) = \sum_j \frac{B_{j(x)} B_{j(y)}}{v} - C. F_{xy} = C_2 \text{ (say) ,}$$

$$\text{and } [d] S. P_{xy} (\text{error}) = T. S. P_{xy} - S. P_{xy} (\text{treat.} + \text{error}) \\ = T. S. P_{xy} - (C_1 + C_2) = C \text{ (say)}$$

(iv) Arrange the *S. S.* and *S. P.* for treatment and error only in the following tabular form—

Source of variation	<i>D. F.</i>	<i>S. S._x</i>	<i>S. S._y</i>	<i>S. P._{xy}</i>
Treatment	$v-1=v_1$	A_1	B_1	C_1
Error	$(r-1)(v-1)=v$	A	B	C
Totals= (Treat.+Error)	v_1+v	(A_1+A)	(B_1+B)	(C_1+C)

Note:—The remaining steps are the same as in the case of C. R. D. except the formulae for S. E. of the difference between the two treatment means (i th and j th). Here, the S. E. of the difference between the two treatment means is given by $\sqrt{V'E \left[\frac{2}{r} + \frac{(\bar{x}_i - \bar{x}_j)^2}{A} \right]}$

Obviously, the S. E. of the difference will be different for different pairs of treatment-means. For most of the practical purposes, it is sufficient to use the average standard error to compare the treatment means. It is given by the formula—

$$\text{Average S. E. of difference} = \sqrt{\frac{2V'E}{r} \left[1 + \frac{1}{v_1} \cdot \frac{A_1}{A} \right]}$$

(3) Case of L. S. D.

• Suppose we have a ' $K \times K$ ' Latin Square given below. Whose A, B, C, D, \dots, K are ' K ' treatments, ' y ' denotes the yield and ' x ' the concomitant variate. $R_{1(y)}$ and $R_{1(x)}$ stand for the 1st row totals for the variates ' y ' and ' x ' respectively and similar meanings are attached with $C_{j(y)}$ and $C_{j(x)}$ for j th column. The symbols $T_{A(y)}$, $T_{A(x)}$ etc denote the treatment totals for the treatment A etc.

Col. ⁿ Row	1 Y, X	2 Y, X	K Y, X	Totals ΣR
1	A y_{11}, x_{11}	B y_{21}, x_{21}	D y_{K1}, x_{K1}	$R_{1(y)}, R_{1(x)}$
2	C y_{12}, x_{12}	A y_{22}, x_{22}	B y_{K2}, x_{K2}	$R_{2(y)}, R_{2(x)}$
⋮	⋮	⋮	⋮	⋮	⋮
K	K y_{1K}, x_{1K}	D y_{2K}, x_{2K}	C y_{KK}, x_{KK}	$R_{K(y)}, R_{K(x)}$
Totals $=C$	$C_{1(y)}, C_{1(x)}$	$C_{2(y)}, C_{2(x)}$	$C_{K(y)}, C_{K(x)}$	$G_{(y)}, G_{(x)}$
Treat.	A	B	K	
Totals	$T_{A(y)},$ $T_{A(x)}$	$T_{B(y)},$ $T_{B(x)}$	$T_{K(y)}$ $T_{K(x)}$	

Further Steps are as follows—

(i) Compute the $S. S_x$ for treatments, rows, columns and error by the following formulae—

$$(a) T. S. S_x = \sum_i \sum_j x_{ij}^2 - C. F_x, \text{ where } C. F_x = \frac{G_{(x)}^2}{N} \text{ and } N = K^2$$

$$(b) S. S_x (\text{treat.}) = \sum \frac{T^2 A_{(x)}}{K} - C. F_x = A_1 (\text{say})$$

$$(c) S. S_x (\text{rows}) = \sum_i \frac{R_{i(x)}^2}{K} - C. F_x = A_2 (\text{say})$$

$$(d) S. S_x (\text{columns}) = \sum_j \frac{C_{j(x)}^2}{K} - C. F_x = A_3 (\text{say})$$

$$(e) S. S_x (\text{error}) = T. S. S_x - S. S_x (\text{treat.} + \text{rows} + \text{columns}) \\ = T. S. S_x - (A_1 + A_2 + A_3) = A (\text{say})$$

(ii) Compute the $S. S_y$ for treatments, rows, columns and error in the following way—

$$(a) T. S. S_y = \sum_i \sum_j y_{ij}^2 - C. F_y, \text{ where } C. F_y = \frac{G_{(y)}^2}{N}$$

$$(b) S. S_y (\text{treat.}) = \sum \frac{T^2 A_{(y)}}{K} - C. F_y = B_1 (\text{say})$$

$$(c) S. S_y (\text{rows}) = \sum_i \frac{R_{i(y)}^2}{K} - C. F_y = B_2 (\text{say})$$

$$(d) S. S_y (\text{columns}) = \sum_j \frac{C_{j(y)}^2}{K} - C. F_y = B_3 (\text{say})$$

$$(e) S. S_y (\text{error}) = T. S. S_y - S. S_y (\text{treat.} + \text{rows} + \text{columns}) \\ = T. S. S_y - (B_1 + B_2 + B_3) = B (\text{say})$$

(iii) Compute the $S. P_{xy}$ for treatments, rows, columns and error by the following formulae—

$$(a) T. S. P_{xy} = \sum_i \sum_j x_{ij} y_{ij} - C. F_{xy}, \text{ where } C. F_{xy} = \frac{G_{(x)} G_{(y)}}{N}$$

$$(b) S. P_{xy} (\text{treat.}) = \sum \frac{T A_{(x)}}{K} \frac{T A_{(y)}}{K} - C. F_{xy} = C_1 (\text{say})$$

$$(c) S. P_{xy} (\text{rows}) = \sum_i \frac{R_{i(x)} R_{i(y)}}{K} - C. F_{xy} = C_2 (\text{say})$$

$$(d) S. P_{xy} (\text{columns}) = \sum_j \frac{C_{j(x)} C_{j(y)}}{K} - C. F_{xy} = C_3 (\text{say})$$

$$(e) S. P_{xy} (\text{error}) = T. S. P_{xy} - S. P_{xy} (\text{treat.} + \text{rows} + \text{columns}) \\ = T. S. P_{xy} - (C_1 + C_2 + C_3) = C (\text{say})$$

(iv) Arrange the *S. S.* and *S. P.* for treatments and error only in the following tabular form—

Source of Variance	<i>D. F.</i>	<i>S. S_x</i>	<i>S. S_y</i>	<i>S. P_{xy}</i>
Treat.	$K-1=v_1$	A_1	B_1	C_1
Error	$(K-1)(K-2)=v$	A	B	C
Totals= (Treatment + Error)	v_1+v	(A_1+A)	(B_1+B)	(C_1+C)

Note—The remaining steps are the same as in the case of *C. R. D.* except the formula for *S. E.* of the difference between the two treatment-means (*i*th, *j*th): Here, the *S. E.* of the difference between the two treatment-means is given by the formula—

$$S. E. \text{ of the difference} = \sqrt{V' E \left[\frac{2}{K} + \frac{(\bar{x}_i - \bar{x}_j)^2}{A} \right]}$$

Obviously, the *S. E.* of the difference for different pairs of treatment-means will be of different magnitudes. But, for most of the practical purposes, it is sufficient to use the average *S. E.* to compare the treatment-means. It is given by the formula—

$$\text{average } S. E. \text{ of difference} = \sqrt{\frac{2V' E}{K} \left(1 + \frac{A_1}{v_1 A} \right)}$$

Exp. No. (1) In a feeding experiment, the rabbits were fed with six types of diets *A, B, C, D, E* and *F*. The following table gives the gains in weight (*y* grams) and the quantity intake (*x* caloric units)—

<i>A</i>		<i>B</i>		<i>C</i>		<i>D</i>		<i>E</i>		<i>F</i>	
<i>x</i>	<i>y</i>	<i>x</i>	<i>y</i>	<i>x</i>	<i>y</i>	<i>x</i>	<i>y</i>	<i>x</i>	<i>y</i>	<i>x</i>	<i>y</i>
10	3.0	7	4.2	13	2.3	6	4.0	9	3.3	12	4.2
8	3.5	6	4.8	10	2.5	9	3.1	14	1.7	8	4.6
15	2.0	5	5.1	12	2.3	7	3.5	16	1.5	6	5.5
11	3.2	9	3.5	7	3.7	10	2.9	8	4.1	11	3.8
9	3.8	8	3.9	9	2.7	5	5.0	11	2.9	14	3.4

Test, whether the gains are materially affected by the quantity of food in take and six types of diets differ significantly ?

Solution—

H₀ : The diets do not differ significantly.

To compute the S , S_x , S_y and S_{xy} , we prepare the following table—

Diets Rep.	A		B		C		D		E		F	
	x	y	x	y	x	y	x	y	x	y	x	y
1.	10	3.0	7	4.2	13	2.3	6	4.0	9	3.3	12	4.2
2.	8	3.5	6	4.8	10	2.5	9	3.1	14	1.7	8	4.6
3.	15	2.0	5	5.1	12	2.3	7	3.5	16	1.5	6	5.5
4.	11	3.2	9	3.5	7	3.7	10	2.9	8	4.1	11	3.8
5.	9	3.8	8	3.9	9	2.7	5	5.0	11	2.9	14	3.4
Totals $T(x), (Ty)$	53	15.5	35	21.5	51	13.5	37	18.5	58	13.5	51	21.5

$$\Sigma x = 285, \quad \Sigma y = 104$$

S. S. for x .

$$C. F_x = \frac{G^2_{(x)}}{N} = \frac{(285)^2}{30} = 270 \times 7.5$$

$$T. S. S_x = \sum_i \sum_j x_{ij}^2 - C. F_x = 2959 - 2707.5 = 251.5$$

$$\begin{aligned} S. S_x (\text{treat. or diet}) &= \sum_i \frac{T_i^2}{r_i} - C. F_x \\ &= \frac{(53)^2 + (35)^2 + \dots + (51)^2}{5} - 2707.5 \\ &= \frac{13969}{5} - 2707.5 = 2793.8 - 2707.5 \\ &= 86.3 \end{aligned}$$

$$\begin{aligned} S. S_x (\text{error}) &= T. S. S_x - S. S_x \text{ due to treat.} \\ &= 251.5 - 86.3 = 165.2 \end{aligned}$$

S. S. for y :

$$C. F_y = \frac{G^2_{(y)}}{N} = \frac{(104)^2}{30} = 360.53$$

$$\begin{aligned} T. S. S_y &= \sum_i \sum_j y_{ij}^2 - C. F_y = 389.46 - 360.53 \\ &= 28.93 \end{aligned}$$

$$\begin{aligned} S. S_y (\text{treat.}) &= \sum_i \frac{T_i^2}{r_i} - C. F_y \\ &= \frac{(15.5)^2 + (21.5)^2 + \dots + (21.5)^2}{5} - 360.53 = \frac{1871.50}{5} \\ &\quad - 360.53 \\ &= 374.30 - 360.53 = 13.77 \end{aligned}$$

$$\begin{aligned} S. S_y (\text{error}) &= T. S. S_y - S. S_y \text{ due to treat.} \\ &= 28.93 - 13.77 = 15.16 \end{aligned}$$

S. P. for x and y :

$$C. F_{xy} = \frac{G_{(x)} G_{(y)}}{N} = \frac{285 \times 104}{30} = 988$$

$$\begin{aligned} T. S. P_{xy} &= \sum_j \sum_i x_{ij} y_{ij} - C. F_{xy} \\ &= 918.7 - 988 = -69.3 \end{aligned}$$

$$\begin{aligned} S. P_{xy} (\text{treat.}) &= \sum_i \frac{T_{i(x)} T_{i(y)}}{r_i} - C. F_{xy} \\ &= \frac{(53 \times 15.5) + \dots + (51 \times 21.5)}{5} - 988 \\ &= \frac{4826.5}{5} - 988 = 965.3 - 988 = -22.7 \end{aligned}$$

$$\begin{aligned} S. P_{xy} (\text{error}) &= T. S. P_{xy} - S. P_{xy} \text{ due to treat.} \\ &= -69.3 - (-22.7) = -46.6 \end{aligned}$$

Now we prepare the following table for unadjusted *S. S* and *S. P.* for treatment and error only—

Source of Variation	<i>D. F.</i>	<i>S. S_x</i>	<i>S S_y</i>	<i>S. P_{xy}</i>	Reg. coeff (<i>b</i>)
Treat.	5	86.3	13.77	-22.7	—
Error	24	165.2	15.16	-46.6	$\frac{-46.6}{165.2} = -0.2821$
Totals= (<i>T</i> + <i>E</i>)	29	251.5	28.93	69.3	—

To test the significance of the regression coefficient '*b*' in the error line, we compute the statistic—

$$F_{(1,23)} = \frac{(-46.6)^2/165.2}{15.16 - (-46.6)^2/165.2} \times 23$$

$$= \frac{13.1450 \times 23}{15.16 - 13.1450} = \frac{302.335}{2.015} = 150.0421$$

$$\therefore F_{(1,23)} \approx 150, \text{ and } F_{(1,23)} = 4.28$$

.05

Thus we have, cal. *F* > tab. *F* at 5% leading to the conclusion that the quantity of food, in take has a significant effect on the gains in weights of the rabbits and the —ve value of regression coefficient ($b = \frac{-46.6}{165.2} = -0.2821$) indicates the —ve correlation between *x* and *y* on the average.

To test the effect of diets, we compute the corrected (adjusted) *S. S.* for *y* (gains in weight) with the help of the following table —

Source of Variation	Unadj. <i>S. S_y</i>	Adjusting factor	Adj. <i>S. S_y</i> =unadj. <i>S. S_y</i> —adj. factor
Treat.	13.77	—	By substraction 7.8196
Error	15.16	$(-46.6)^2/165.2 = 13.1450$	2.015
Totals (<i>T</i> + <i>E</i>)	28.93	$(-69.3)^2/251.5 = 19.0954$	9.8346

A. V. T. for adjusted $S. S_y$:

Source of Variation	$D. F.$	$\circ S. S_y$	$M. S. S_y$	$F. cal.$	F at	
					5%	1%
Treat.	5	7.8196	1.56392	17.85**	2.64	3.94
Error	23	2.015	0.0876 $= V'E$	—	—	—
Totals	28	9.8346	—	—	—	—

The significant value of F corresponding to the treatments (diets) indicates that the diets are not homogeneous but differ significantly.

Finally to compare the diets, we compute the adjusted mean gains in weight of rabbits due to diets A, B, C, D, E and F and also their $C. D_s$ with the help of the following table—

$$\bar{x} = \frac{\sum x}{n} = 9.5, b = \frac{-46.6}{165.2} = -0.2821$$

Diet	$\bar{x}_i = \frac{T_{i(x)}}{r_i}$	$\bar{x}_i - \bar{x}$	$b(\bar{x}_i - \bar{x})$	$y_i = \frac{T_{i(y)}}{r_i}$	Adjusted mean $y_i' = y_i - b(\bar{x}_i - \bar{x})$
A	10.6	1.1	-0.3103	3.1	3.4103
B	7.0	-2.5	0.7053	4.3	3.5947
C	10.2	0.7	-0.1975	2.7	2.8.75
D	7.4	-2.1	0.5924	3.7	3.1076
E	11.6	2.1	-0.5924	2.7	3.2924
F	10.2	0.7	-0.1975	4.3	4.4975

(i) $S. E.$ of difference between the two adjusted means of F and C .

$$= \sqrt{V'E \left[\frac{1}{r_6} + \frac{(\bar{x}_6 - \bar{x}_3)^2}{A} + \frac{1}{r_3} \right]}$$

where $r_4 = r_3 = 5$

$$V'E = 0.0876$$

(adjusted error variance) $A = 165.2$

$$= \sqrt{0.0876 \left[\frac{1}{5} + \frac{(10.2 - 10.2)^2}{165.2} + \frac{1}{5} \right]}$$

$$= \sqrt{0.0876 \times 0.4} = \sqrt{0.03504} = 0.11872$$

$$(C. D.) = (\underset{5\%}{S. E. \text{ of difference}}) \times \underset{.05}{t_{(23)}} = 0.872 \times 2.069 = 0.3873$$

$$\text{But } \bar{y}_6 - \bar{y}_3 = 4.4975 - 2.8975 = 1.60$$

As the difference between the adjusted average values of F and C is greater than their (C. D.), the diet F differs significantly from C at 5% level of significance.

(ii) $S. E.$ of difference between the two adjusted means of F and A .

$$\therefore S.E. = \sqrt{V'E \left[\frac{1}{r_6} + \frac{(\bar{x}_6 - \bar{x}_1)^2}{A} + \frac{1}{r_1} \right]}$$

$$= \sqrt{0.0876 \left[\frac{1}{5} + \frac{(10.2 - 10.6)^2}{165.2} + \frac{1}{5} \right]} = \sqrt{0.0876 \times 0.401}$$

$$= \sqrt{0.0351276} = 0.1874$$

$$(C. D.) = (\underset{5\%}{S. E. \text{ of difference}}) \times \underset{.05}{t_{(23)}} = 0.1874 \times 2.069 = 0.3877$$

$$\text{But } \bar{y}_6 - \bar{y}_1 = 4.4975 - 3.4103 = 1.0872$$

Evidently, the difference between the adjusted average values of F and A is greater than their (C. D.) at 5% level.

i. e. $1.0872 > 0.3877$, so the diet F differs significantly from A at 5% level.

Similarly, the remaining pairs may be compared after computing the C. D. for each pair.

Inference—The six diets produced the different gains in the weight of rabbits and the diet F is the best of all.

Note—The calculations in the above example may be easier if the deviations be taken from $y = 3.0$ and $x = 11$ for the two variates respectively.

Exp. No. (2) A varietal on cotton was laid out in 5 blocks and 5-varieties of cotton namely *A, B, C, D* and *E* were tested. On each plot the total plant population (denoted by x) and cotton yield (denoted by y in Kgms.) were recorded. The following table gives the data—

Blocks Variety	1		2		3		4		5	
	x	y	x	y	x	y	x	y	x	y
A	27	60	21	63	27	60	21	66	27	57
B	33	70	15	49	27	66	23	66	24	59
C	28	64	25	64	33	64	25	64	25	56
D	24	65	22	64	21	67	19	65	26	65
E	32	65	24	62	28	47	18	61	35	61

Test whether the yields are materially affected by the plant population/plot and the five varieties differ significantly ?

Solution—

Ho : The data is homogeneous with respect to the blocks and the varieties.

For convenience in calculations, we prepare the following table after taking deviations in x from $x=25$ and in y from $y=60$ Kgms. respectively.

Blocks Variety	1		2		3		4		5		Totals	
	x	y	x	y	x	y	x	y	x	y	$T_{(x)}$	$T_{(y)}$
A	2	0	-4	3	2	0	-4	6	2	-3	-2	6
B	8	10	-10	-11	2	6	-2	6	-1	-1	-3	10
C	3	4	0	4	8	4	0	4	0	-4	11	12
D	-1	5	-3	4	-4	7	-6	5	1	5	-13	26
E	7	5	-1	2	-3	13	-7	1	10	1	12	-4
Totals $B_{(x)}, B_{(y)}$	19	24	-18	2	11	4	-19	2	12	-2	5	50
											$G_{(x)}$	$G_{(y)}$

S. S. for x :

$$C. F_x = \frac{G^2_{(x)}}{N} = \frac{(5)^2}{25} = 1$$

$$T. S. S_x = \sum_{i,j} x^2_{ij} - C. F_x = (2)^2 + (8)^2 + \dots + (10)^2 - 1 = 561 - 1 = 560$$

$$S. S_x (\text{blocks}) = \sum_j \frac{B^2_{j(x)}}{5} - C. F_x = \frac{(19)^2 + (-18)^2 + (11)^2 + (-19)^2 + (12)^2}{5} - 1 = 262 \cdot 2 - 1 = 261 \cdot 2$$

$$S. S_x (\text{treat. or variety}) = \sum_i \frac{T^2_{i(x)}}{5} - C. F_x = \frac{(-2)^2 + (-3)^2 + \dots + (12)^2}{5} - 1 = 89 \cdot 4 - 1 = 88 \cdot 4$$

$$S. S_x (\text{error}) = T. S. S_x \text{ due to } (\text{blocks} + \text{treat.}) = 560 - (261 \cdot 2 + 88 \cdot 4) = 560 - 349 \cdot 6 = 210 \cdot 4$$

S. S. for y :

$$C. F_y = \frac{G^2_{(y)}}{N} = \frac{(50)^2}{25} = 100$$

$$T. S. S_y = \sum_{i,j} y^2_{ij} - C. F_y = (10)^2 + \dots + (1)^2 - 100 = 768 - 100 = 668$$

$$S. P_{xy} (\text{blocks}) = \sum_j \frac{B^2_{j(y)}}{5} - C. F_y = \frac{(24)^2 + (2)^2 + \dots + (-4)^2}{5} - 100 = 216 \cdot 8 - 100 = 116 \cdot 8$$

$$S. S_y (\text{treat.}) = \sum_i \frac{T^2_{i(y)}}{5} - C. F_y = \frac{(6)^2 + (10)^2 + \dots + (-4)^2}{5} - 100 = 194 \cdot 4 - 100 = 94 \cdot 4$$

$$S. S_y (\text{error}) = T. S. S_y - S. S_y \text{ due to } (\text{blocks} + \text{treat.}) = 668 - (116 \cdot 8 + 94 \cdot 4) = 668 - 211 \cdot 2 = 456 \cdot 8$$

S. P. for x and y :

$$C. F_{xy} = \frac{G_{(x)} G_{(y)}}{N} = \frac{5 \times 50}{25} = 10$$

$$T. S. P_{xy} = \sum_{i,j} x_{ij} y_{ij} - C. F_{xy} = (2 \times 0) + (8 \times 10) + \dots + (10 \times 1) - 10 = 120 - 10 = 110$$

$$S. P_{xy} (\text{blocks}) = \sum_j \frac{B_{j(x)} B_{j(y)}}{5} - C.F_{xy} = \frac{(19 \times 24) + \dots + (12 \times -2)}{5} - 10$$

$$= 4.4 - 10 = -5.6$$

$$S. P_{xy} (\text{treat.}) = \sum_i \frac{T_{i(x)} T_{i(y)}}{5} - C.F_{xy} = \frac{(-2 \times 6) + \dots + (12 \times -4)}{5} - 10$$

$$= -59.2 - 10 = -69.2$$

$$S. P_{xy} (\text{error}) = T. S. P_{xy} - S. P_{xy} \text{ due to (blocks + treat.)}$$

$$= 110 - (-5.6 - 69.2)$$

$$= 184.8$$

Now we prepare the following table for unadjusted *S. S.* and *S. P.* for treatments and error only—

Source of Variation	<i>D. F</i>	<i>S. S_x</i>	<i>S. S_y</i>	<i>S. P_{xy}</i>	Reg. Coefficient 'b'
Treatments	4	88.4	94.4	-69.2	—
Error	16	210.4	456.4	184.8	$\frac{184.8}{210.4} = 0.8783$
Totals (Treat. + error)	20	298.8	551.2	115.6	—

To test the significance of the regression coefficient 'b' in the error line, we compute the statistic—

$$F_{(1, 15)} = \frac{(184.8)^2 / 210.4}{456.8 - (184.8)^2 / 210.4} \times 15$$

$$= \frac{162.0148}{294.4852} \times 15 = \frac{2434.7220}{294.4852} = 8.2677$$

$$\therefore F_{(1, 15)} = 8.2677, \text{ and } F_{(1, 15)} = 4.54$$

05

The +ve value of regression coefficient 'b' indicates that there is +ve correlation between *x* (plant population per plot) and *y* (yield of cotton) and the significant value of *F* corresponding to regression coefficient shows that the plant population/plot has a significant effect on the yield of cotton.

To test the effect of treatments (varieties), we compute the adjusted $S. S_y$ with the help of the following table—

Source of Variation	Unadj. $S. S_y$	Adjusting factor	Adj. $S. S_y = \text{Unadj. } S. S_y - \text{adj. factor}$
Treat.	94.4	—	By subtraction 211.9914
Error	456.8	$(184.8)^2/210.4 = 162.3148$	$456.8 - 162.3148 = 294.4852$
Totals= ($T+E$)	551.2	$(115.6)^2/298.8 = 44.7234$	$551.2 - 44.7234 = 506.4766$

A. V. T. for adjusted $S. S_y$ —

Source of variation	$D. F$	$S. S_y$	$M. S. S_y$	F cal	F at	
					5%	1%
Treatments	4	211.9914	52.99785	2.6995	3.06	4.89
Error	15	294.4852	$19.6323 = V'_E$	—	—	—
Totals	19	506.4766	—	—	—	—

The calculated value of F corresponding to the treatments (varieties) comes out to be insignificant and so the varieties A, B, C, D and E are homogeneous.

Now we need to compute the adusted mean yields (\bar{y}_i'), only which are shown in the following table—

$$\bar{x} = \frac{5}{25} + 25 = 25.2, b = 0.8783,$$

Variety	$\bar{x}_i = \frac{T_{i(x)}}{r}$	$(\bar{x}_i - \bar{x})$	$b(\bar{x}_i - \bar{x})$	$\bar{y}_i = \frac{T_{i(y)}}{r}$	$\bar{y}_i' = \bar{y}_i - b(\bar{x}_i - \bar{x})$
A	24.6	-0.6	-0.5270	61.2	61.7170
B	24.4	-0.8	-0.7026	62.0	62.7026
C	27.2	2.0	1.7566	62.4	60.6434
D	22.4	-2.8	-2.4592	65.2	67.6592
E	27.4	2.2	1.9323	59.2	57.2677

Performance of yields—

Variety :	D	B	A	C	E
adj. mean yield (in kgms).	67.6592,	62.7026,	61.7270,	60.6534,	57.2677

The variety D is the best of all since it gives the maximum adj. mean yield of cotton.

Inference : The varieties of cotton do not differ significantly and the variety D is the best of all.

Exp No. (3) An experiment on cotton was carried out in a 5×5 L. S. to test the homogeneity of five varieties of cotton, namely A, B, C, D and E. On each plot the yield (y kgms) and the total number of plants were recorded as follows—

	A	B	C	D	E
x	27	21	27	21	27
y	60	63	60	66	57
	B	C	D	E	A
x	33	15	27	23	24
y	70	49	66	66	59
	C	D	E	A	B
x	28	25	33	25	25
y	64	64	64	64	56
	D	E	A	B	C
x	24	22	21	19	26
y	65	67	67	65	65
	E	A	B	C	D
x	32	24	28	18	35
y	65	62	47	61	61

Test whether the yield (y) is materially affected by the total number of plants per plot and the varieties of cotton differ significantly?

Solution :—

Ho : The data is homogeneous.

Taking the deviations from $x=25$ and $y=60$, we prepare the following table—

Rows	No. (x) and yield(y)	COLUMNS					Totals $R_{(x)}$ $R_{(y)}$
		1	2	3	4	5	
1	x y	A 2 0	B -4 3	C 2 0	D -4 6	E 2 —	—2
2	x y	B 8 10	C -10 —11	D 2 6	E -2 6	A -1 —1	—3 10
3	x y	C 3 4	D 0 4	E 8 4	A 0 4	B 0 —4	11 12
4	x y	D -1 5	E -3 4	A -4 7	B -6 5	C 1 5	—13 26
5	x y	E 7 5	A -1 2	B 3 —13	C -7 1	D 10 1	12 —4
Totals $C_{(x)}$ $C_{(y)}$	x y	19 24	—18 2	11	—19 22	12 —2	$G_{(x)}=5$ $G_{(y)}=50$
Treat. $T_{(x)}$ $T_{(y)}$	x y	A -4 12	B 1 1	C -11 —1	D 7 22	E 12 16	

S. S. for x.

$$C. F._x = \frac{G_{(x)}^2}{N} = \frac{(5)^2}{25} = 1$$

$$T. S. S._x = \sum_{i,j} x_{ij}^2 - C. F._x = (2)^2(2) + \dots + (10)^2 + \dots - 1 = 561$$

$$-1 = 560$$

$$S. S._x \text{ (Rows)} = \sum \frac{R_{i(x)}^2}{r} - C. F._x$$

$$\frac{(-2)^2 + (-3)^2 + \dots + (12)^2}{5} - 1 = 89.4 - 1 = 88.4$$

$$\begin{aligned}
 S. S. _x (\text{columns}) &= \sum_j \frac{C_{j(x)}^2}{5} - C. F. _x \\
 &= \frac{(19)^2 + (-18)^2 + \dots + (12)^2}{5} - 1 = 262.2 - 1 = 261.2
 \end{aligned}$$

$$\begin{aligned}
 S. S. _x (\text{treat. or variety}) &= \sum_i \frac{T_{i(x)}^2}{5} - C. F. _x \\
 &= \frac{(-4)^2 + (1)^2 + (-11)^2 + (7)^2 + (12)^2}{5} - 1 = 66.2 - 1 = 65.2
 \end{aligned}$$

$$\begin{aligned}
 S. S. _x (\text{error}) &= T. S. S. _x - S. S. _x \text{ due to (Rows + cols + treats.)} \\
 &= 560 - (88.4 + 261.2 + 65.2) = 145.2
 \end{aligned}$$

S. S. for y.

$$C. F. _y = \frac{G_{(y)}^2}{N} = \frac{(50)^2}{25} = 100$$

$$T. S. S. _y = \sum_{i,j} y_{ij}^2 - C. F. _y = (0)^2 + (10)^2 + \dots + (1)^2 - 100 = 768 - 100 = 668$$

$$\begin{aligned}
 S. S. _y (\text{Rows}) &= \sum_i \frac{R_{i(y)}^2}{5} - C. F. _y \\
 &= \frac{(6)^2 + (10)^2 + \dots + (-4)^2}{5} - 100 = 194.4 - 100 = 94.4
 \end{aligned}$$

$$\begin{aligned}
 S. S. _y (\text{columns}) &= \sum_j \frac{C_{j(y)}^2}{5} - C. F. _y \\
 &= \frac{(24)^2 + (2)^2 + \dots + (-2)^2}{5} - 100 = 216.8 - 100 = 116.8
 \end{aligned}$$

$$\begin{aligned}
 S. S. _y (\text{treat.}) &= \sum_i \frac{T_{i(y)}^2}{5} - C. F. _y \\
 &= \frac{(12)^2 + (1)^2 + (-1)^2 + (22)^2 + (16)^2}{5} - 100 \\
 &= 177.2 - 100 = 77.2
 \end{aligned}$$

$$\begin{aligned}
 S. S. _y (\text{error}) &= T. S. S. _y - S. S. _y \text{ due to (Rows + cols. + treats.)} \\
 &= 668 - (94.4 + 116.8 + 77.2) = 379.6
 \end{aligned}$$

S. P. for x & y,

$$C. F. _{xy} = \frac{G_{(x)} G_{(y)}}{5} = \frac{5 \times 25}{25} = 10$$

$$\begin{aligned}
 T. S. P. _{xy} &= \sum_{i,j} x_{ij} y_{ij} - C. F. _{xy} \\
 &= (2 \times 0) + (8 \times 10) + \dots + (10 \times 1) - 10 \\
 &= 120 - 10 = 110
 \end{aligned}$$

$$\begin{aligned}
 S. P_{xy} \text{ (Rows)} &= \sum_i \frac{R_{i(x)} R_{i(y)}}{5} - C. F_{xy} \\
 &= \frac{(-2 \times 6) + \dots + (12 \times -4)}{5} - 10 \\
 &= -59.2 - 10 = -69.2
 \end{aligned}$$

$$\begin{aligned}
 S. P_{xy} \text{ (columns)} &= \sum_j \frac{C_{j(x)} C_{j(y)}}{5} - C. F_{xy} \\
 &= \frac{(19 \times 24) + \dots + (12 \times -2)}{5} - 10 \\
 &= -4.4 - 10 = -5.6
 \end{aligned}$$

$$\begin{aligned}
 S. P_{xy} \text{ (treats)} &= \sum_i \frac{T_{i(x)} T_{i(y)}}{5} - C. F_{xy} \\
 &= \frac{(-4 \times 12) + (1 \times 1) + (-11 \times -1) + (7 \times 22) + (12 \times 16)}{5} - 10 \\
 &= 62.0 - 10 = 52.0
 \end{aligned}$$

$$\begin{aligned}
 S. P_{xy} \text{ (error)} &= T. S. P_{xy} - S. P_{xy} \text{ due to (Rows + cols. + treats.)} \\
 &= 110 - (-69.2 - 5.6 + 52.0) = 132.8
 \end{aligned}$$

Now we prepare the following table for unadjusted $S. S.$ and $S. P.$ for treatment and error only—

Source of variation	D. F.	$S. S._x$	$S. S._y$	$S. P_{xy}$	Reg. coeff. 'b'
Treat.	4	65.2	77.2	52.0	—
Error	12	145.2	379.6	132.8	$\frac{132.8}{145.2} = 0.9146$
Totals ($T+E$)	16	210.4	456.8	184.8	

To test the significance of the regression-coefficient 'b' in the error line, we compute the statistic—

$$\begin{aligned}
 F(1, 11) &= \frac{(132.8)^2/145.2}{379.6 - (132.8)^2/145.2} \times 11 \\
 &= \frac{121.4590}{258.1410} \times 11 \\
 &= \frac{1336.0490}{258.1410} = 5.1757
 \end{aligned}$$

$$\therefore F(1, 11) = 5.1757 \text{ and } F_{05}(1, 11) = 4.84$$

The +ve value of regression coefficient 'b' indicates that there is +ve correlation between x (no. of plants/plot) and y (yield of cotton) and the significant value of F corresponding to the regression coefficient shows that the plant population/plot has a significant effect on the yield of cotton.

To test the effect of treatments (varieties), we compute the adjusted $S. S_y$ with the help of the following table—

Source of variation	Unadj. $S. S_y$	Adjusting factor	Adj. $S. S_y$ = Unadj. $S. S_y$ - adj. factor
Treat.	65.2	—	By Subtraction 36.3442
Error	145.2	$\frac{(132.8)^2}{145.2} = 121.4590$	$379.6 - 121.4590 = 258.1410$
Totals (T+E)	210.4	$\frac{(184.8)^2}{210.4} = 162.3148$	$456.8 - 162.3148 = 294.4852$

A. V. T. for adjusted $S. S_y$ —

Source of variation	D. F.	$S. S_y$	$M. S. S_y$	F cal.	F at	
					5%	1%
Treat.	4	36.3442	9.08605	2.58	5.93	14.45
Error	11	258.1410	$\frac{23.4674}{= V'_E}$	—	—	—
Totals	15	294.4852	—	—	—	—

Calculated F corresponding to the treatments (varieties) comes out to be insignificant and so the varieties A, B, C, D and E are homogeneous.

Now we need to compute the adjusted mean yields (y'_i) only which are shown in the following table—

$$\bar{x} = 25.2, \quad b = 0.9146$$

Variety	$\bar{x}_i = \frac{T_{i(x)}}{r}$	$(\bar{x}_i - \bar{x})$	$b(\bar{x}_i - \bar{x})$	$y_i = \frac{T_{i(y)}}{r}$	$y_i - b(\bar{x}_i - \bar{x}) = y'_i$
A	24.2	-1.0	-0.9146	62.4	63.3146
B	25.2	0	0	60.2	60.2
C	22.8	-2.4	-2.1950	59.8	61.9950
D	26.4	1.2	1.0975	64.4	63.3035
E	27.4		2.0121	63.2	61.1879

Performance of Yields—

Variety : A D C E B
 Adj. mean yield : 63.3146, 63.3025, 61.9950, 61.1879, 60.2
 (in Kgms.)

Inference—The varieties do not differ significantly and the variety A is the best of all.

Exercise V

Q. 1. What is cocomitant variate and how it can be used for improving the precision of experimental results? Explain with examples and indicate the method of Statistical analysis of results?

(M. Sc. Ag. Agra, 1956)

Q. 2. Discuss the uses of analysis of covariance in experimentation?

In a randomized block experiment with ' r ' replications, the degrees of freedom, sum of squares and sum of products for treatments and error are denoted as follows—

	$D. F.$	$S(X^2)$	$S(XY)$	$S(Y^2)$
Treatments	p	A_1	B_1	C_1
Error :	q	A_2	B_2	C_2

Outline briefly the procedure for carrying out the adjusted analysis of variance for ' y '?

(M. Sc. Ag. Agra, 1962)

Q. 3. In a $R. B. D.$ with 4 replications and 16 treatments the following results were recorded :-

Source :	$D. F.$	$S. S_r$ (unadj.)	$S. S_y$ (unadj.)	$S. P_{ry}$ (unadj.)
Treatment :	—	10777.5	13.5	112.5
Error :	—	15698.2	36.0	248.0

Test whether the yield (y) is materially affected by the plant-population (r) per plot and the six treatments differ significantly?

Ans. $F_{(1,44)} = 5.37$, $F_{(5,44)} = 3.43$

CHAPTER VI

Analysis of Incomplete Observations.

Or

MISSING PLOT TECHNIQUE

In field experimentation, whatever care the experimenter may take in designing and conducting the experiment, the yields of some plots may not be obtained correctly. They may be depredated by cattles, wild animals and birds or affected seriously by some pest, disease or flood etc. Sometimes, it may be the case that the yields of some plots may be affected by some factor which does not affect the others. Such plots affected from extraneous sources would not provide unbiased comparisons and hence their yield have to be omitted from the analysis of the data. In the case either the yields are missing or have been omitted, the data is incomplete and its statistical analysis is somewhat more complex than that of the complete data. Here, we shall deal with the case with one missing observation only for *R. B. D.* and *L. S. D.*

(i) Analysis of *R. B. D.* with one missing unit :

Suppose, we have got '*v*' treatments each replicated '*r*' times in the original plan and the yield of (i, j) th plot be missing. The yields can be arranged in the following tabular form—

Blocks Treat	1	2	3	<i>j</i>	<i>r</i>	Totals
1	y_{11}	y_{12}	y_{13}	y_{1j}	y_{1r}	T_1
2	y_{21}	y_{22}	y_{23}	y_{2j}	y_{2r}	T_2
3	y_{31}	y_{32}	y_{33}	y_{3j}	y_{3r}	T_3
⋮	⋮	⋮	⋮	⋮	⋮	⋮
<i>i</i>	y_{i1}	y_{i2}	y_{i3}	X	y_{ir}	T_i
⋮	⋮	⋮	⋮	⋮	⋮	⋮
<i>v</i>	y_{v1}	y_{v2}	y_{v3}	y_{vj}	y_{vr}	T_v
Totals	B_1	B_2	B_3	B	B_r	G

Where X denotes the missing value, T is the total of i th treatment and B is the total of j th block (replicate) containing the missing value.

G is the grand total of $(vr-1)$ values.

There are two methods of analysing the above data—

(a) **Bartlett's Method**—Bartlett first made the use of analysis of covariance-technique in analysing the incomplete data. For a single missing value, the method consists in giving the value zero to the missing yield and then supposing an imaginary variate ' x ' which assumes the value zero for all the plots except the missing plot where it takes the value -1 and finally applying yield as dependent variate.

Standard Error (S. E.)—If the treatments come out to be significant at α % level of significance, then the S. E. of the difference between the two treatments, none of which has the missing value, is

$\sqrt{\frac{2V'E}{r}}$, where $V'E$ is the adjusted error-variance and that of between the two treatment means, one of which has the missing value, is $\sqrt{\frac{V'E}{r} \left[2 + \frac{v}{(r-1)(v-1)} \right]}$

Mean of the i th treatment (for which the value is missing)

The mean of the missing value treatment is

$\frac{1}{r} \left[T + \frac{(rB+vT-G)}{(r-1)(v-1)} \right]$ and the rest procedure is the same as in the case of a complete data.

Exp. No. (1) In the following table, we are given the number of pods/plant. Three plants are selected at random from each of the four blocks but the yield of the plant of block number 3 with variety v_2 could not be recorded. Carry out the analysis of the following incomplete data and state your conclusions ?

Variety		V_1	V_2	V_3	Totals
Blocks	B_1	15	17	15	47
	B_2	12	19	18	49
	B_3	10	X	17	$27=B$
	B_4	15	20	16	51
	Totals	52	56	66	$174=G$
					$=T$

Solution—

• **Ho :** The data is homogeneous with respect to the varieties and the blocks. • • •

The Statistical analysis will be done in the following tabular form—

Variety Blocks \	V_1	V_2	V_3	Totals
B_1	15 (225) 0	17 (289) 0	15 (225) 0	47 (739) 0
B_2	12 (144) 0	19 (361) 0	18 (324) 0	49 (829) 0
B_3	10 (100) 0	0 (0) -1	17 (289) 0	27 = B (389) -1
B_4	15 (225) 0	20 (400) 0	16 (256) 0	51 (881) 0
Totals	52 (694) 0	56 = T -1 (1050)	66 (1094) 0	174 = G -1 (2838)

Now we compute $C. F_s$ as follows—

$$C. F_x = \frac{(\sum x)^2}{N} = \frac{(1)^2}{12} = 0.083$$

$$C. F_y = \frac{(\sum y)^2}{N} = \frac{(174)^2}{12} = \frac{30276}{12} \\ = 2523.0$$

$$C. F_{xy} = \frac{\sum x \cdot \sum y}{N} = \frac{(-1)(174)}{12} = -14.5$$

The figures within brackets indicate the squares of the yields ' y_s '

S. S. for x :—

$$T. S. S_x = \sum_i \sum_j x_{ij}^2 - C. F_x = 1 - 0.083 = 0.917$$

$$\begin{aligned} S. S_x \text{ due to blocks} &= \sum_j \frac{B_j^2}{3} - C. F_x \\ &= \frac{(0)^2 + (0)^2 + (-1)^2 + (0)^2}{3} - 0.083 \\ &= 0.333 - 0.083 = 0.250 \end{aligned}$$

$$\begin{aligned} S. S_x \text{ due to varieties} &= \sum_i \frac{T_i^2}{4} - C. F_x \\ &= \frac{(0)^2 + (-1)^2 + (0)^2}{4} - 0.083 \\ &= 0.250 - 0.083 = 0.167 \end{aligned}$$

$$\begin{aligned} S. S_x \text{ due to error} &= T. S. S_x - S. S_x \text{ due to (blocks} \\ &\quad \text{+ varieties)} \\ &= 0.917 - (0.250 + 0.167) \\ &= 0.917 - 0.417 = 0.500 \end{aligned}$$

S. S. for y :

$$\begin{aligned} T. S. S_y &= \sum_i \sum_j y_{ij}^2 - C. F_y = (15)^2 + (17)^2 + \dots + (16)^2 - 2523.0 \\ &= 2838 - 2523 = 315.0 \end{aligned}$$

$$\begin{aligned} S. S_y \text{ (blocks)} &= \sum_j \frac{B_j^2}{3} - C. F_y = \frac{(47)^2 + (49)^2 + (27)^2 + (51)^2}{3} - 2523.0 \\ &= \frac{2209 + 2401 + 729 + 2601}{3} - 2523.0 \\ &= \frac{7940}{3} - 2523.0 \\ &= 2646.67 - 2523.0 \\ &= 123.67 \end{aligned}$$

$$\begin{aligned} S. S_y \text{ (varieties)} &= \sum_i \frac{T_i^2}{4} - C. F_y = \frac{(52)^2 + (56)^2 + (66)^2}{4} - 2523.0 \\ &= \frac{2704 + 3136 + 4356}{4} - 2523.0 \\ &= \frac{10196}{4} - 2523.0 = 2549 - 2523 \\ &= 26.00 \end{aligned}$$

$$S. S_y (\text{error}) = T. S. S_y - S. S_y \text{ due to (blocks + varieties)} \\ = 315.00 - (123.67 + 26.00) = 165.33$$

S. P. (Sum of products) for x, y :

$$T. S. P_{xy} = \sum_i \sum_j x_{ij} y_{ij} - C. E_{xy} = 0 - (-14.5) = 14.5$$

$$S. P_{xy} (\text{blocks}) = \frac{(\sum x \sum y)_{B_1} + (\sum x \sum y)_{B_2} + (\sum x \sum y)_{B_3} + (\sum x \sum y)_{B_4}}{3} - C. F_{xy}$$

$$= \frac{(0 \times 47) + (0 \times 49) + (-1)(27) + (0 \times 51)}{3} - C. F_{xy}$$

$$= -9.0 + 14.5 = 5.5$$

$$S. P_{xy} (\text{varieties}) = \frac{(\sum x \sum y)_{V_1} + (\sum x \sum y)_{V_2} + (\sum x \sum y)_{V_3}}{4} - C. F_{xy}$$

$$= \frac{(0 \times 52) + (-1)(56) + (0 \times 66)}{4} - (-14.5)$$

$$= -14.0 + 14.5 = 0.5$$

$$S. P_{xy} (\text{error}) = T. S. P_{xy} - S. P_{xy} \text{ due to (blocks + varieties)} \\ = 14.5 - (5.5 + 0.5) = 8.5$$

Now we frame the table for unadjusted $S. S.$ & $S. P.$ —

Source of variation	D.F.	S. S. _x	S. S. _y	S. P. _{xy}
Blocks	3	0.250	123.67	5.5
Varieties	2	0.167	26.00	0.5
Error	6	0.500	165.33	8.5
Totals	11	0.917	315.00	14.5

The adjusting factors and adjusted S. S. for 'y' are computed in the following table for treatment: (varieties) and error only—

Source of variation	S. S.	S. S.	S. P.	Adjusting factors	Adj S S	D. F.	M. S. S.	F. cal.	F. 05
Varieties	0.167	26.00	0.50	—	69.88—20.80 =49.08	2	24.54	5.89*	5.79
Error	0.500	165.30	8.50	144.50	165.3—144.5 =20.80	5	4.16 =V'E		13.27
Total = (Var. + Error)	0.667	191.30	9.0	(9) ² /0.667 =121.44	191.3—121.44 =69.88	7	—	—	—

The value of cal. F corresponding to the varieties comes out to be significant at 5% level and so we require the computations of S. E., S_s and C. D., S_s for the variety-means.

(i) S. E. of the difference between the two variety-means (except of V_2)

$$= \sqrt{\frac{2V'E}{r}} = \sqrt{\frac{2 \times 4.16}{4}} = \sqrt{2.08} = 1.442$$

(where no value is missing)

(ii) *S. E* of the difference between the two variety-means (V_2 and any other)

$$= \sqrt{\frac{V_1 E}{r} \left[2 + \frac{v}{(r-1)(v-1)} \right]}$$

$$= \sqrt{\frac{4 \cdot 16}{4} \left[2 + \frac{3}{3 \times 2} \right]} = \sqrt{1 \cdot 04 \times 2 \cdot 5} = \sqrt{2 \cdot 60} = 1 \cdot 6124$$

(where one variety has one missing value)

(i) (C. D.) = (*S. E.* of difference) \times *t*(5)

$$\begin{array}{ccc} 5\% & & \cdot 05 \\ = 1 \cdot 4422 \times 2 \cdot 571 & = & 3 \cdot 7079 \approx 3 \cdot 71 \end{array}$$

(ii) (C. D.) = (*S. E.* of difference) \times *t*(5)

$$\begin{array}{ccc} 5\% & & \cdot 05 \\ = 1 \cdot 6124 \times 2 \cdot 571 & = & 4 \cdot 1455 \approx 4 \cdot 15 \end{array}$$

The variety-means are calculated as follows—

$$\text{mean of } V_1 = \frac{52}{4} = 13 \cdot 0$$

$$\text{mean of } V_3 = \frac{66}{4} = 16 \cdot 5$$

$$\text{mean of } V_2 = \frac{1}{r} \left[T + \frac{(rB + vT - G)}{(r-1)(v-1)} \right]$$

(since v_2 has one missing value)

$$= \frac{1}{4} \left[56 + \frac{(4 \times 27 + 3 \times 56 - 174)}{3 \times 2} \right]$$

$$= \frac{1}{4} [56 + 17] = 18 \cdot 25$$

Now we arrange the variety-means in descending order of their magnitudes—

Variety :	V_2	V_3	V_1
Mean :	18.25	16.60	13.00

Inference : The varieties V_1 , V_2 & V_3 differ significantly in yielding-capacity at 5% level. We also note that V_2 has the maximum yielding-capacity followed by V_3 but their difference is not significant. The least yield has been observed in the case of variety V_1 which does not differ significantly from V_3 but differs significantly from V_2 .

(b) **Substitution-Method**—The method of substitution consists of the following steps—

(i) Estimate the missing value by using the formula—

$$X = \frac{(rB + vT - G)}{(r-1)(v-1)}$$

(ii) Substitute this value of X for the missing plot yield and compute the $S. S.$ as if no value is missing.

(iii) Subtract the quantity $\frac{(B + vT - G)^2}{v(v-1)(r-1)^2}$ from the sum of squares due to treatments to get the adjusted sum of squares for the treatments.

(iv) Reduce the $d. f.$ for total and error by one, as one missing value is estimated from the data.

(i) Prepare the analysis of variance table and test the significance of the treatment-effect.

(vi) Calculate the $S. E.$ of the difference between the two treatment means, none of which contains the missing value by the relation—

$$S. E. \text{ of difference} = \sqrt{\frac{2V_E}{r}} \text{ and that of the difference}$$

between the two treatment means, one of which contains the missing value, by the formula—

$$S. E. \text{ of difference} = \sqrt{V_E \left[\frac{1}{r-1} + \frac{1}{r-\frac{1}{2}} \right]}, \quad \text{if the}$$

treatments show significant effects.

Exp. No. (2.—Estimate the missing value in example no. (1) and test the significance of the difference between the means of the varieties ?

Solution :

Ho : The data is homogeneous with respect to varieties and blocks.

First of all, we estimate the missing value in the data by the formula—

$$\begin{aligned} x &= \frac{(rB + vT - G)}{(r-1)(v-1)} \\ &= \frac{(4 \times 27 + 3 \times 56 - 174)}{3 \times 2} = \frac{102}{6} = 17.0 \end{aligned}$$

Substituting this estimated value for the missing yield, we take the deviations from $y=16$ for our convenience and prepare the following table--

Variety Block	V ₁	V ₂	V ₃	Totals = B	(Totals) ² = B
B ₁	-1 (1)	-1 (1)	-1 (1)	-1 (3)	1
B ₂	-4 (6)	3 (9)	2 (4)	1 (29)	1
B ₃	-6 (36)	1 (1)	1 (1)	-4 (38)	16
B ₄	-1 (1)	4 (16)	0 (0)	3 (17)	9
Totals = T	-12 (54)	9 (27)	2 (6)	-1 = G (87)	1
(Totals) ² = T	144	81	4	1	
Variety	V ₁	V ₂	V ₃		
Mean	13	18.25	16.5		

$$C. F. = \frac{G^2}{N} = \frac{(-1)^2}{12} = 0.083$$

$$T. S. S. = \sum_{i,j}^34 y_{ij}^2 - C. F. = 87 - 0.083 = 86.917$$

$$S. S. (blobs) = \sum_j \frac{B_j^2}{3} - C. F. = \frac{1+1+16+9}{3} - 0.083 \\ = 9 - 0.083 = 8.917$$

$$S. S. (varieties) = \sum_i \frac{T_i^2}{4} - C. F. \\ = \frac{144+81+4}{4} - 0.083 = 57.250 - 0.083 \\ = 57.167$$

$$S. S. (error) = T. S. S. - S. S. \text{ due to (blocks + varieties)} \\ = 86.917 - (8.917 + 57.167) = 86.917 - 66.084 \\ = 20.833$$

$$\text{The adjusting factor for variety } S. S. = \frac{(B + vT - G)^2}{v(v-1)(r-1)^2} \\ \text{or } \frac{[B - (v-1)x]^2}{v(v-1)} \\ = \frac{[27 - 2 \times 17]^2}{3 \times 2} = \frac{49}{6} \\ = 8.167$$

$$\therefore S. S. (varieties) \text{ adjusted} = S. S. (varieties) \text{ unadj.} - \text{adj. factor} \\ = 57.167 - 8.167 = 49.00$$

Now we arrive at the following *A. V. T.*—

Source of variation	<i>D. F.</i>	<i>S. S.</i>	<i>M. S. S.</i>	<i>F. cal.</i>	<i>F. tab. at</i>	
					5%	1%
Blocks	3	8.917	2.9723	$\frac{4.1666}{2.9723}$ = 1.39	9.01	28.24
Varieties (adj.)	2	49.00	24.5	$\frac{24.5}{4.1666}$ = 5.87*	5.79	13.27
Error	5	20.833	$\frac{4.1666}{= V_E}$	—	—	—

The calculated value of F corresponding to the varieties comes out to be significant at 5% level and it necessitates the further computations of $S. E_s$ and $C. D_s$ as given below—

(i) $S. E$ of the difference between the two variety-means

(except of V_2)

$$\sqrt{\frac{2V_E}{r}} = \sqrt{\frac{2 \times 4 \cdot 1666}{4}} = \sqrt{2 \cdot 0833}$$

(where no variety has missing unit)

(ii) $S. E$ of difference between the two variety-means

(V_2 and any other)

$$= \sqrt{\frac{V_E}{r} \left[\frac{1}{r-1} + \frac{1}{r-\frac{1}{2}} \right]} \quad \sqrt{4 \cdot 1666 \left[\frac{1}{3} + \frac{1}{3 \cdot 5} \right]}$$

$$= \sqrt{4 \cdot 1666 \times 0 \cdot 6190} = \sqrt{2 \cdot 5791254} = 1 \cdot 6059$$

(where one variety contains one missing unit)

(i) $(C. D.) = (S. E. \text{ of difference}) \times t$ (5)

5%

05

$$= 1 \cdot 4433 \times 2 \cdot 571 = 3 \cdot 7118 \approx 3 \cdot 71$$

(ii) $(C. D.) = (S. E. \text{ of difference}) \times t$ (5)

5%

05

$$= 1 \cdot 6059 \times 2 \cdot 571 = 4 \cdot 1288 \approx 4 \cdot 13$$

Now we arrange the varieties according to their performance—

Variety :	V_2	V_3	V_1
Mean :	18·25	16·50	13·00

Inference—Here we draw the same conclusions as given in exp. (1) on page (99).

(i) Analysis of Latin Square Design with one missing unit :

Suppose we have got a ' $K \times K$ ' L. S. D. with (i , j th) plot yield missing and

R is the total of the row containing the missing value,

C is the total of the columns containing the missing value,

T is the treatment total containing the missing value and

G is the grand total of (K^2-1) observations.

The analysis of this incomplete design can be carried out by using the analysis of 'covariance-technique', but the 'substitution-method' is simpler and more rapid than the former. The steps involved in the 'substitution-method' are—

(i) Estimate the missing value by the formula

$$X = \frac{K(R+T+C)-2G}{(K-1)(K-2)},$$

(ii) Substitute this value of X for the missing observation and calculate the S. S. in the usual way,

(iii) Subtract the quantity $\frac{[(K-1)T+R+C-G]^2}{[(K-1)(K-2)]^2}$ from the treatment S. S.,

(iv) Reduce the total and error d. f. by one since one missing value is estimated in the data.

(v) Prepare the analysis of variance table to test the homogeneity of the data.

(vi) Compute the S. E. of the difference between the two treatment means, none of which is attached with the missing value by using the relation—

$$\text{S. E. of the difference} = \sqrt{\frac{2V_E}{K}}, \text{ and that of the difference}$$

one of which is attached with the missing value, by the formula—

$$\text{S. E. of the difference} = \sqrt{2V_E \left(\frac{1}{K-1} + \frac{1}{K-2} \right)}, \text{ if the}$$

treatments show significant effect.

Exp. No. (3) : Estimate the missing value in the following L. S. D. and carry out the analysis of variance to test the significance of difference between the treatment-means ?

B	D	A	C
249	245	249	244
A	C	D	B
254	249	240	252
D	B	C	A
245	254	250	257
C	A	B	D
251	261	×	246

Solution :

Ho : The data is homogeneous.

The missing value is estimated by the formula—

$$x = \frac{K(R+T+C)-2G}{(K-1)(K-2)},$$

where $K=4$, $R=758$, $C=739$, $T=755$ and $G=3746$ from the data.

$$\begin{aligned}\hat{x} &= \frac{4(758+755+739)-2 \times 3746}{3 \times 2} = \frac{9008-7492}{6} = \frac{1516}{6} = 252.66 \\ &\approx 252.7\end{aligned}$$

Using the estimated value 252.7 for the missing unit and taking the deviations from $y=250$, we prepare the following table to compute the sum of squares—

Col. ⁿ Row	1	2	3	4	Totals = R	(Totals = R) ²
1	B -1 (1)	D -5 (25)	A -1 (1)	C -6 (36)	-13 (63)	169
2	A 4 (16)	C -1 (1)	D -10 (100)	B 2 (4)	-5 (25)	25
3	D -5 (25)	B 4 (16)	C 0 (0)	A 7 (49)	6 (90)	36
4	C 1 (1)	A 11 (121)	B 2.7 (7.29)	D -4 (16)	10.7 (145.29)	114.49
Totals=C	-1 (43)	9 (163)	-8.3 (108.29)	-1 (105)	-1.3=G (419.29)	1.69
(Totals=C) ²	1	81	68.89	1	1.69	
Treat.	A 21 (441)	B 7.7 (59.29)	C -6 (36)	D -24 (576)		
Mean	255.25	251.925	248.5	256.0		

$$C. F. = \frac{G^2}{N} = \frac{1.69}{16} = 0.105625 \approx 0.1056$$

$$T. S. S. = \sum \sum y_{ij}^2 - C. F. = 419.29 - 0.1056 = 419.1844$$

$$\begin{aligned} S. S. \text{ columns} &= \sum_j \frac{C_j^2}{4} - C. F. = \frac{1+81+68.89+1}{4} - 0.1056 \\ &= \frac{151.89}{4} - 0.1056 \\ &= 37.9725 - 0.1056 = 37.8669 \end{aligned}$$

$$\begin{aligned} S. S. (\text{Row}) &= \sum_i \frac{R_i^2}{4} - C. F. = \frac{169+25+36+114.49}{4} - 0.1056 \\ &= \frac{344.49}{4} - 0.1056 = 86.1225 - \\ &\quad - 0.1056 \\ &= 86.0169 \end{aligned}$$

$$\begin{aligned} S. S. (\text{treat.}) &= \frac{T^2 A + T^2 B + T^2 C + T^2 D}{4} - C. F. \\ &= \frac{441 + 59.29 + 36 + 576}{4} - 0.1056 \\ &= 278.0725 - 0.1056 = 277.9669 \end{aligned}$$

$$\begin{aligned} S. S. (\text{error}) &= T. S. S. - S. S. \text{ due to (rows + columns + treat.)} \\ &= 419.1844 - (86.0169 + 37.8669 + 277.9669) \\ &= 419.1844 - 401.8507 = 17.3337 \end{aligned}$$

$$\begin{aligned} \text{The adjusting factor for treatment } S. S. &= \frac{[(K-1)T + R + C - G]^2}{[(K-1)(K-2)]^2} \\ &= \frac{[3 \times 755 + 758 + 739 - 3746]^2}{[3 \times 2]^2} = \frac{(16)^2}{(6)^2} = \frac{256}{36} \\ &= 7.1111 \end{aligned}$$

$$\begin{aligned} \therefore \text{adjusted } S. S. (\text{treat.}) &= \text{unadj. } S. S. (\text{treat.}) - \text{adj. factor} \\ &= 277.9669 - 7.1111 = 270.8558 \end{aligned}$$

Now we arrive at the following *A. V. T.*—

Source of Variation	<i>D. F.</i>	<i>S. S_y</i>	<i>M. S. S_y</i>	<i>F. cal.</i>	<i>F at</i>	
					5%	1%
Rows	3	86.0169	28.6723	8.27*	5.41	12.06
Columns	3	37.8669	12.6223	3.64	„	„
Treat.(adj.)	3	270.8558	90.2853	26.05**	„	„
Error	5	17.3337	3.4665 $= V_E$	—	—	—
Totals	14	—	—	—	—	—

The calculated value of *F* corresponding to rows comes out to be significant at 5% level and that of treatments is highly significant. The insignificant value of *F* corresponding to columns indicates that the design has no improvement over the *R. B. D.* in this example. In order to test the significance of the difference between the two treatment-means, we compute below the *S. E_s* and *C. D_s* —

(i) The *S. E.* of the difference between the two treatment-means

(except of *B*) is given by the formula $\sqrt{\frac{2V_E}{K}}$

$$= \sqrt{\frac{2 \times 3.4665}{4}} = \sqrt{1.73325} = 1.3165$$

(where no treatment has missing value)

The *S. E.* of the difference between the two treatment-means (*B* and any other) is given by the formula—

$$\begin{aligned} & \sqrt{V_E \left(\frac{1}{K-1} + \frac{1}{K-\frac{2}{3}} \right)} \\ &= \sqrt{3.4665 \left(\frac{1}{3} + \frac{3}{10} \right)} = \sqrt{3.4665 \times 0.6333} = \sqrt{2.19533445} \\ &= 1.4816 \end{aligned}$$

(where one treatment has one missing unit)

$$(i) (C. D.)_{.05} = (S. E. \text{ of difference}) \times t_{.05} \quad (5)$$

$$= 1.3165 \times 2.571 = 3.384 \approx 3.385$$

$$\text{and } (C. D.)_{.01} = (S. E. \text{ of difference}) \times t_{.01} \quad (5)$$

$$= 1.3165 \times 4.032 = 5.3081 \approx 5.308$$

$$(ii) (C. D.)_{.05} = (S. E. \text{ of difference}) \times t_{.05} \quad (5)$$

$$= 1.4816 \times 2.571 = 3.8092 \approx 3.809$$

$$\text{and } (C. D.)_{.01} = (S. E. \text{ of difference}) \times t_{.01} \quad (5)$$

$$= 1.4816 \times 4.032 = 5.9738 \approx 5.974$$

Now we arrange the treatment-means in the decreasing order of their magnitudes—

(i) For 5% level,	Treatment : D	A	B	C
	Mean : 256.0	255.25	251.925	248.5

(ii) For 1% level,	Treatment : D	A	B	C
	Mean : 256.0	255.25	251.925	248.5

Inference—The maximum yield has been recorded in the case of treatment *D* and minimum by *C*. The difference between *D* and *C* is significant at both 5% and 1% levels but the difference between *D* and *B* is significant at 5% and not at 1%.

Exp. No. (4)—Compare the advantages and disadvantages of the *R. B.* and *L. S.* designs in field trials. In a trial of five varieties of wheat, *A, B, C, D* and *E* laid out in a *L. S.*, the following yields (in oz. per plot) were obtained—

	B	E	C	A	D	Totals
	90	80	134	—	92	396
	E	D	B	C	A	
	85	84	70	141	82	462
	C	A	D	B	E	
	111	90	87	84	69	441
	A	C	E	D	B	
	81	125	85	76	72	439
	D	B	A	E	C	
	82	60	94	85	88	409
Totals	449	439	470	386	403	2147

The yield of plot under variety *A* in the first row was lost on account of damage by cattle. Calculate the missing value and the adjustment to the treatment sum of squares in the analysis of variance of the completed data. State the formula for the *S. E* or comparison of two varieties involving variety *A* ?

(M. Sc. Ag. Agra, 1958)

Solution—

For the first part of the question, see theory.

Ho : The data is homogeneous.

The missing value is estimated by the formula—

$$\hat{x} = \frac{K(R+C+T)-2G}{(K-1)(K-2)}$$

where $K=5$, $R=396$, $C=386$, $T=347$ and $G=2147$ from the data.

$$\begin{aligned}\hat{x} &= \frac{5(396+386+347)-2 \times 2147}{4 \times 3} \\ &= \frac{5645-4294}{12} = \frac{1351}{12} = 112.5833 \approx 112.6\end{aligned}$$

Substituting this estimated value 112.6 for the missing yield and taking the deviations from $y=90$ (for convenience in calculation), we get the data in the following tabular form —

Col ⁿ \ Row	1	2	3	4	5	Totals=R	(Totals=R) ²
1	B 0 (0)	E -10 (100)	C (1936)	A 22.6 (510.76)	D 2 (4)	58.6 (2550.76)	3433.96
2	E -5 (25)	D -6 (36)	B -20 (400)	C 51 (2601)	A -8 (64)	12 (3126)	144
3	C 21 (441)	A 0 (0)	D -3 (9)	B -6 (36)	E -21 (441)	-9 (927)	81
4	A -9 (81)	C 35 (1225)	E -5 (25)	D -14 (196)	B -18 (324)	-11 (1851)	121
5	D -8 (64)	B -30 (900)	A 4 (16)	E -5 (25)	C -2 (4)	0 -41 (1009)	1681
Totals=C	-1	-11	20	48.6	-47	96=G	92.16
(Totals=C) ²	(611)	(2261)	(2386)	(3368.76)	(837)	(9463.76)	
Variety	A 9.6 (92.16)	B -74 (5476)	C 149 (22201)	D -29 (841)	E -46 (2116)		
Mean	91.92	75.2	119.8	84.2	80.8		

$$C. F. = \frac{G^2}{N} = \frac{92 \cdot 16}{5} = 3 \cdot 63611$$

$$T. S. S. = \sum_i \sum_j y_{ij}^2 - C. F. = 9463 \cdot 76 - 3 \cdot 6864 = 9460 \cdot 0736$$

$$\begin{aligned} S. S. (\text{rows}) &= \sum_i \frac{R_i^2}{5} - C. F. \\ &= \frac{3433 \cdot 96 + 144 + 81 + 121 + 1681}{5} - 3 \cdot 6864 \\ &= 1092 \cdot 1920 - 3 \cdot 6864 = 1088 \cdot 5056 \end{aligned}$$

$$\begin{aligned} S. S. (\text{columns}) &= \sum_j \frac{C_j^2}{5} - C. F. \\ &= \frac{1 + 121 + 400 + 2351 \cdot 96 + 2209}{5} - 3 \cdot 6864 \\ &= 1018 \cdot 5920 - 3 \cdot 6864 = 1014 \cdot 9056 \end{aligned}$$

$$\begin{aligned} S. S. (\text{varieties}) &= \sum \frac{T^2}{5} - C. F. \\ &= \frac{92 \cdot 16 + 5476 + 22201 + 841 + 2116}{5} - 3 \cdot 6864 \\ &= 6145 \cdot 2320 - 3 \cdot 6864 = 6141 \cdot 5456 \end{aligned}$$

$$\begin{aligned} S. S. (\text{error}) &= T. S. S. - S. S. \text{ due to (rows + columns + varieties)} \\ &= 9460 \cdot 0736 - (1088 \cdot 5056 + 1014 \cdot 9056 + 6141 \cdot 5456) \\ &= 9460 \cdot 0736 - 8244 \cdot 6568 = 1215 \cdot 1168 \end{aligned}$$

The adjusting factor for variety (treatment) S. S. is given by the formula—

$$\begin{aligned} \text{adjusting factor} &= \frac{[(K-1)T + R + C - G]^2}{[(K-1)(K-2)]^2} \\ &= \frac{[4 \times 347 + 396 + 386 - 2147]^2}{[4 \times 3]^2} \\ &= \frac{(23)^2}{(12)^2} = 3 \cdot 6736 \end{aligned}$$

$$\begin{aligned} \therefore \text{adjusted } S. S. (\text{varieties}) &= \text{unadjusted } S. S. (\text{varieties}) - \text{adj. factor} \\ &= 6141 \cdot 5456 - 3 \cdot 6736 \\ &= 6137 \cdot 8720 \end{aligned}$$

Now we arrive at the following A. V. T.—

Source of variation	D.F.	S. T.	M. S. S.	F cal.	F at	
					5%	1%
Rows	4	1088.5056	272.1264	2.46	3.36	5.67
Columns	4	1014.9056	253.7264	2.29	„	„
Varieties (adj.)	4	6137.8720	1535.4680	13.89**	„	„
Error	11	1215.1168	110.4652 $=V_E$	—	„	—
Totals	—	—	—	—	—	—

The calculated value of F corresponding to the varieties comes out to be highly significant and those of rows and columns both are insignificant showing that the design has no improvement over C. R. D. in this example.

The formula for the S. E. of the difference between the variety-means involving the variety A (which has a missing unit) is given by—

$$S. E. \text{ of difference} = \sqrt{V_E \left(\frac{1}{K-1} + \frac{1}{K-\frac{2}{3}} \right)},$$

where V_E is the error variance (error mean square)

$$\begin{aligned}
 &= \sqrt{110.4652 \left(\frac{1}{4} + \frac{1}{5-\frac{2}{3}} \right)} \\
 &= \sqrt{110.4652 \times \frac{25}{12}} = \sqrt{53.10826923} \\
 &= 7.2875
 \end{aligned}$$

Result : The missing value = 112 approx.,

The varieties differ significantly at 1% level of significance. The S. E. of the difference between the two variety-means involving A, is given by $\sqrt{V_E \left[\frac{1}{K-1} + \frac{1}{K-\frac{2}{3}} \right]}$ and comes out as 7.2875.

EXERCISE VI

Q. (1): (a) What are the methods of estimating a missing value in R. B. D. ?

(b) The table gives the results of a randomized block experiment in which the observation indicated by X is missing. Estimate the missing value and analyse the data ?

Treat.	Replication						Totals
	I	II	III	IV	V	VI	
A :	17	29	25	17	33	23	144
B :	19	23	X	15	23	19	99
C :	33	35	29	25	37	27	186
Totals	69	87	54	57	93	69	429

Ans. (b) $\times = 19.2$. $V_E = 7.438 V_t (\text{adj}) = 174.90$

Q. (2): The yields/plot (in kgms) are recorded below on the basis of an experiment performed on four varieties of barley each with 6 replications. But the yield for the variety A could not be recorded in the sixth plot as it was lost on account of damage by cattle. Estimate the missing yield and carryout the analysis of variance with adjusted sum of squares for variety and state your conclusions ?

Variety	Replications						Totals
	1	2	3	4	5	6	
A :	15	17	15	17	19	a	83
B :	21	19	15	19	17	17	108
C :	19	17	17	21	19	17	110
D :	21	23	19	25	22	17	127
Totals	76	76	66	82	77	51	428

Ans. $a = 14$, $V_E = 2.61905 V_t (\text{adj}) = 23.3611$

Q. (3) A feeding experiment was conducted on a dozen of cows of four different breeds which were grouped into 3 such that each group consists of 4 cows of different breeds. The cows in 3 groups were subjected to three types of diets (treatments) A, B and C and the increase in milk-yields (ounces/cow) after a week were recorded as follows—

Diets/breeds	I	II	III	IV
A :	8.0	26.0	18.0	21.0
B :	22.0	25.0	24.0	36.0
C :	11.0	27.0	13.0	X

were \times denotes the missing yield for the cow of 4th breed which was supplied the diet C.

(i) Estimate the missing yield and carryout the analysis of variance to test the significance of difference between the 4 breeds and three 3 diets ?

(ii) State the formulae for comparison of two diets involving the diet C and excluding the diet C ?

(iii) Write down the adjusting factor to get the adjusted treatment sum of squares ?

Ans : (i) $X=25$, $V_E=22.9667$ $V_t(\text{adj.})=84.50$

$$(ii) \sqrt{V_E \left[\frac{1}{r-1} + \frac{1}{r-\frac{1}{2}} \right]} \text{ and } \sqrt{\frac{2V_E}{r}}$$

for S. E. of difference of treatment means.

$$(iii) [B-(v-1)X]^2/[v(v-1)]$$

Q. (4) . Estimate the missing value (denoted by X) in the following 4×4 L. S. D. and carryout the analysis of variance to test the homogeneity of the data and also state the formulae for S. E. of difference between the two treat-means involving the treatment which has one missing value and excluding it ?

C	B	A	D
25	23	20	20
A	D	C	B.
19	19	21	18
B	A	D	C
19	X	17	20
D	C	B	A
17	20	20	15

where letters stand for treatments.

Ans. $X=17.0$, $V_E=1.0$,

$$V_t(\text{adj.})=10.1852,$$

$$\sqrt{V_E \left[\frac{1}{K-1} + \frac{1}{K-\frac{2}{3}} \right]} \text{ and } \sqrt{\frac{2V_E}{K}}$$

Q. 5. In a trial of five varieties of wheat *A, B, C, D* and *E*, laid out in a Latin Square, the following yields (in oz./plot) were obtained. But the yield for the variety *A* in the second row and 5th column was lost on account of damage by animals.

(i) Calculate the missing yield and the adjusting factor to the treatment (variety) *S. S.* in the analysis of variance of the completed data ?

(ii) Analyse the data and interpret your results obtained ?

	A	B	C	D	E	Totals
	50	70	70	80	90	360
B		C	D	E	A	
	70	90	80	80	—	320
C		D	E	A	B	
	60	50	90	80	90	370
D		E	A	B	C	
	50	60	80	50	70	310
E		A	B	C	D	
	80	90	50	85	60	365
Totals	10	360	370	375	310	1725

Ans. (i) $X=100$, adj. factor = 76.5625

(ii) $V_F = 216.3636$ V_F (adj.) = 270.8594

Q. 6. We are giving Below some results of a 5×5 L. S. experiment conducted on 25 cows of 5 different breeds and of 5 lactation periods fed with 5 types of rations *A, B, C, D* and *E*, for a month. The data for increase in milk-yields (in gms.) were observed after the month but the increase in milk-yield for the cow of 4th breed and 1st lactation period supplied with diet *A*, could not be recorded.

Row (lactation) total containing the missing-yield = 396 gms.

Columns (breed) „ „ „ „ = 386 „

Total under diet (treatment) *A* = 347 „

Grand sum = 2147 „

T. S. S. = 9460.0736

Source of Variation	<i>D. F.</i>	<i>S. S</i>	<i>M. S. S.</i>	<i>F</i>
Rows (lactation periods)	—	—	272·1264	—
Columns (breeds)	—	—	253·7264	—
Diets (Treats.) adj.	—	—	—	—
Error	11	1215·1168	—	—
Totals	23			

[a] Estimate the missing increase in milk-yield ?

[b] Calculate the adjusting factor to treatment sum of squares and obtain the adjusted treatment *S. S.* ?

[c] Also complete the *A. V. T.* and state your conclusions regarding the homogeneity of breeds, lactation periods and the different diets ?

Ans. [a] $X=112\cdot6$ approx

[b] adj. factor= $3\cdot6736$,
adj. T_r *S. S.*= $6137\cdot8720$

[c] $F=2\cdot46$ for rows (lactation periods)

$F=2\cdot29$ for columns (breeds)

$F=13\cdot89$ for treatments

CHAPTER VII

Factorial Experiments

Concept—In the foregoing experiments performed either in *C. R. D.*, *R. B. D.* or *L. S. D.*, we were concerned only with the variation in a simple factor like different varieties and manures, different supply-rates of the same manure and cultural treatments etc. In field-experimentation, very often the situations arise when we have to test the variation in a no. of factors simultaneously. For example we may be interested in selecting the best variety of all the available ones of a certain crop and rate of nitrogen supply for a newly acquired land. In order to achieve the object, we may first compare the varieties in absence of nitrogen and adopting the best variety as shown by the experiment proceed to select the best rate of nitrogen supply by performing a second experiment. The conclusions drawn as above are valid only when the effect exerted by any one of the two factors is independent of the other but it is not always true. In most of the cases, the response of the first factor varies according to the levels of the second factor (i. e. the two factors interact each other). For example, a higher level of irrigation is essential to secure an adequate response of yield when a heavy dose of nitrogen has been applied to a certain crop. Thus the use of the above discussed scheme is limited. Another point against this approach is that the precision of the two experiments are different and hence the results cannot be compared.

The most informative and efficient approach when several factors are under study, is to compare all the possible combinations of the different levels of all the factors simultaneously in the same experiment. This approach is known as the *Factorial Concept of experimentation*. As it compares all the factors in the one and the

same experiment with equal precision, there is a greater scope for this type of experimentation in comparison to the traditional methods and it renders the comparisons of the results. It also saves a great deal of time and the experimental material. The main advantage of the factorial scheme is that it can be used to study the simultaneous variation among several factors whether they are independent or interact each other and provides a way to test the significance of the interaction between two or more factors.

Definitions and Symbolic representations—Before proceeding to the analysis of the factorial experiments, we set forth some definitions and give the symbolic representations to be used.

(i) Case of 2^2 factorial experiment :

In order to have a clear conception of the terms to be defined, we consider an experiment of wheat with two factors each at two levels. These are nitrogen, none (n_0) versus 22 kgms/Hect. (n_1) and superphosphate, none (p_0) versus 22 kgms/Hect. (p_1). This experiment is called a 2×2 or 2^2 factorial-experiment. Let us suppose that the mean yields (in maunds/acre) of the four possible treatment combinations are as given below—

S. No.	1	2	3	4
Treat. comb. :	n_0p_0	n_0p_1	n_1p_0	n_1p_1
Mean yields :	26	30	28	34

These yields can be put in the following tabular-form—

Super phosphate	Nitrogen		Response ($n_1 - n_0$)
	n_0	n_1	
p_0	26	28	2
p_1		34	4
Response ($p_1 - p_0$)	4	6	—

From the above table, we see that—

(i) The application of super phosphate has increased the yield by 4 maunds/acre in the absence of nitrogen and 6 maunds/acre in the presence of nitrogen. These are called the *Simple-effects* of super phosphate. The average effect $\left(\frac{4+6}{2}=5 \text{ maunds/acre}\right)$ of its application is called the *Main effect* of Super-phosphate.

(ii) The application of nitrogen has increased the yield of wheat by 2 maunds/acre in the absence of super-phosphate and 4 maunds/acre in the presence of super phosphate. These are called the *Simple effects* of nitrogen. The average effect $\left(\frac{2+4}{2}=3 \text{ maunds/acre}\right)$ of its application is called the *Main effect* of nitrogen.

(iii) The 2 simple effects of nitrogen are not the same, this fact indicates that the two factors are not independent. If they were so, then the increase in yields due to the application of nitrogen should have been the same at the two different levels of the super-phosphate. Similarly, the two simple effects of super phosphate are not the same leading to the conclusion that the two factors are not independent but interact each other. *The measure of the extent to which the two factors interact is given either by half of the difference between the simple effect of nitrogen in presence of the super phosphate and that of in the absence of super phosphate* (i. e. $\frac{4-2}{2}=1 \text{ maund/acre}$) or by half of the difference between the simple effect of super phosphate in the presence of nitrogen and that of in the absence of nitrogen (i. e. $\frac{6-4}{2}=1 \text{ maund/acre}$).

For simplicity, the first level of any factor in treatment combinations is signified by its absence and the suffix of the second level is dropped. The treatment combination consisting first level of all the factors is denoted by (1). Now the four treatment-combinations n_0p_0 , n_0p_1 , n_1p_0 & n_1p_1 can be written as—

(1), p, n & np respectively. The symbol (p) stands for the total yield of the plots receiving the treatment combination n_0p_1 or p. Similar meanings are attached with the alike symbols.

Computation of main effects and interactions : For computing a main effect or interaction, first we write down its corresponding expression $\frac{1}{2r} (n \pm 1) \cdot (p \pm 1)$; where 'r' is the no. of replications and any bracket contains either +ve or -ve sign according to the absence or presence of the corresponding capital letters in the symbol for the main effect or the interaction to be computed. Then expanding the expression algebraically and substituting the yields for the treatment combinations. Thus—

$$\begin{aligned}
 N &= \frac{1}{2r} (n-1)(n+1) \\
 &= \frac{1}{2r} [np + n - p_1 - (1)] \\
 &= \frac{1}{2} \left[\frac{np}{r} + \frac{(n)}{r} - \frac{(p)}{r} - \frac{(1)}{r} \right] \\
 &= \frac{1}{2} [34 + 28 - 30 - 26] = 3
 \end{aligned}$$

Symbolic representation : It is customary to denote the factors, their main effects and interactions by *Capital letters* and the different levels and their combinations by small letters. In the present example.

N → denotes the factor nitrogen and the main effect of nitrogen,

P → denotes the factor Super phosphate and the main effect of Super phosphate,

NP or PN → denotes the interaction of the two factors Nitrogen and super phosphate. It is called *two factor-interaction* or *first-order interaction*.

n_0 → denotes the first level of nitrogen, (none)

n_1 → „ „ second „ „ ,
(22 kgm/Hect.)

p → „ „ first „ „ , (none)

p_1 → „ „ second „ „ ,
(22 kgm/Hect)

$n_0p_0 \rightarrow$	„ „ combination of nitrogen and Super phosphate both at first level,
$n_0p_1 \rightarrow$	„ „ combination of nitrogen at first level and phosphate at second level,
$n_1p_0 \rightarrow$	„ „ combination of nitrogen at second level and phosphate at first level,
$n_1p_1 \rightarrow$	„ „ combination of nitrogen and phosphate both at second level.

$$\begin{aligned}
 P &= \frac{1}{2r} \left[(p-1)(n+1) \right] \\
 &= \frac{1}{2r} \left[np + p - n - (1) \right] \\
 &= \frac{1}{2} \left[\frac{(np)}{r} + \frac{(p)}{r} - \frac{(n)}{r} - \frac{(1)}{r} \right] \\
 &= \frac{1}{2r} \left[34 + 30 - 28 - 26 \right] = 5 \\
 NP &= \frac{1}{2r} \left[(p-1)(n-1) \right] \\
 &= \frac{1}{2r} \left[np - p - n + (1) \right] \\
 &= \frac{1}{2} \left[\frac{(np)}{r} - \frac{(p)}{r} - \frac{(n)}{r} + \frac{(1)}{r} \right] \\
 &= \frac{1}{2} \left[34 - 30 - 28 + 26 \right] = 1
 \end{aligned}$$

The symbol $[N] = [(n-1)(p+)]$ stand for the total factorial effect of nitrogen and the similar meanings for others.

Statistical analysis : The factorial experiments are performed either in C. R. D., R. B. D. or L. S. D. and so their analysis under factorial scheme remains the same except that the treatment-sum of square is split up into its components each with 1' d. f. In the present example, the component sum of squares are computed as given below—

Treat. component	D. F.	S. S.
N	1	$\frac{[(n-1)(p+1)]^2}{2^2 r} = [N]^2/2^2 r$
P	1	$[(n+1)(p-1)]^2/2^2 r = [P]^2/2^2 r$
NP	1	$[(n-1)(p-1)]^2/2^2 r = [NP]^2/2^2 r$
Totals	3	S. S. (treat.)

By yate's method : It is a simple and rapid method giving us a convenient way of computing the treatment component sum of square. The computational work is carried out in the following tabular form—

Treat. comp. in standard order	Total yield	I	II	Total fact. effect	S. S.	du c to
(1)	(1)	(1)+(n)	(1)+(n) +(p)+(np)	G	$\frac{G^2}{2^2 r} = C.F.$	—
n	(n)	(p)+(np)	(n)-(1) +(np)-(p)	[N]	$[N]^2/2^2 r$	N
p	(p)	(n)-(1)	(p)+(np) -(1)-(n)	[P]	$[P]^2/2^2 r$	P
np	(np)	(np)-(p)	(np)-(p) -(n)+(1)	[NP]	$[NP]^2/2^2 r$	NP

The first two figures in column I are totals of the two pairs and the last two figures are differences of the same pairs in the column of total yield. (Where the first value of a pair is subtracted from the second). The column II has been obtained by the similar operations on the pairs of the column I.

Exp. (1) An experiment was planned to study the effect of sulphate potash and super phosphate on yield of potatoes. All the combinations of 2 levels of super phosphate [0 cent (p_0) and 5 cent (p_1)/acre] and two levels of sulphate of potash [0 cent (k_0) and 5 cent (k_1)/acre] were studied in a randomized block design with 4 replications for each. The following yield (lbs per plot = $\frac{1}{70}$ acre) were obtained—

Block	(1)	k	p	kp
I	23	25	22	38
	p	(1)	k	kp
II	40	26	36	38
	(1)	k	pk	p
III	29	20	30	20
	kp	k	p	(1)
IV	34	31	24	28

Analyse the data and give your conclusions ?

Solutions : Taking deviations from $y=29$, we prepare the following table for computations of the S. S. due to treatments and blocks—

Treat. comb. \ Block	I	II	III	IV	Totals = T	(T totals = T)
(1)	-6 (36)	-3 (9)	0 (0)	-1 (1)	-10 (46)	100
k	-4 (16)	7 (49)	-9 (81)	2 (4)	-4 (150)	16
p	-7 (49)	11 (121)	-9 (81)	-5 (25)	-10 (276)	100
kp	9 (81)	9 (81)	1 (1)	5 (25)	24 (188)	576
Totals = B	-8 (182)	24 (260)	-17 (163)	1 (55)	0 = G (660)	0
(Totals = B) ²	64	576	289	1	0	

Ho : The data is homogeneous with respect to the blocks and the treatments.

$$C. F. = \frac{G^2}{N} = \frac{(0)^2}{16} = 0$$

$$T. S. S. = \sum_i \sum_j y_{ij}^2 - C. F. = 660 - 0 = 660$$

$$S. S. (\text{blocks}) = \sum_j \frac{B_j^2}{4} - C. F. \\ = \frac{64 + 576 + 289 + 1}{4} - 0 = \frac{930}{4} = 232.50$$

$$S. S. (\text{treat}) = \sum_i \frac{T_i^2}{4} - C. F. = \frac{100 + 16 + 100 + 576}{4} - 0 \\ = \frac{792}{4} = 198.0$$

$$S. S. (\text{error}) = T. S. S. - S. S. \text{ due to (blocks + treat.)} \\ = 660 - (232.50 + 198.0) = 660 - 430.5 = 229.50$$

The treatment component sum of squares are obtained by Yates's Method—

Treatment combination	Total yield	I	II	Total factorial effect	S. S.	due to
(1)	-10	-14	0	G	$(0)^2/16=0$ =C F.	—
k	-4	14	40	[K]	$(40)^2/16=100$	K
p	-10	6	28	[P]	$(28)^2/16=49$	P
kp	24	34	28	[KP]	$(28)^2/16=49$ Total =198	KP

Now we prepare the following *A. V. T.*—

Source of variation	D F.	S. S.	M. S. S.	F cal.	F at	
					5%	1%
Blocks	3	232.5	77.5	3.04	3.86	6.99
Treat.	3	198.0	66.0	2.59	3.86	6.99
K	1	100	100	3.92	5.12	10.56
P	1 = 3	49 = 198	49	1.92	„	„
KP	1	49	49	1.92	„	„
Error	9	229.5	25.5	—	—	—
Totals	15	660.0	—	—	—	—

The calculated values of F corresponding to the blocks, treatments and treatment-components come out to be insignificant for all.

Inference—There is no significant difference between the blocks and the treatments *i. e.* the data is homogeneous with respect to the blocks and the treatments.

S. E. : The S. E. of any effect or interaction in the case of a 2^n factorial experiment is
$$= \sqrt{\frac{V_E}{r \cdot 2^{n-2}}}$$

For $n=2$, the S. E.
$$= \sqrt{\frac{V_E}{r}}$$

and for $n=3$, the S. E.
$$= \sqrt{\frac{V_E}{2r}}$$

Case of 2^3 factorial experiment—In case of 2^3 factorial experiment (*i e* 3 factors each at two levels) a_0, a_1, b_0, b_1 and c_0, c_1 respectively. Then the eight possible treatment combinations in standard order will be given as (1), a, b, ab, c, ac, bc and abc . A 2^3 factorial experiment can be conducted either in C. R. D., R. B. D. or L. S. D. and its analysis remains the same except that the treatment S. S. is split up into its components each with 1 d. f. By the straight forward extension of the rules given in 2^2 factorial experiment, we have treatment-component S. S. as given below—

Treatment component	D.F.	S. S.
A	1	$[(a-1)(b+1)(c+1)]^2/2^3.r = [A]^2/2^3.r$
B	1	$[(a+1)(b-1)(c+1)]^2/2^3.r = [B]^2/2^3.r$
AB	1	$[(a-1)(b-1)(c+1)]^2/2^3.r = [AB]^2/2^3.r$
C	1	$[(a+1)(b+1)(c-1)]^2/2^3.r = [C]^2/2^3.r$
AC	1	$[(a-1)(b+1)(c-1)]^2/2^3.r = [AC]^2/2^3.r$
BC	1	$[(a+1)(b-1)(c-1)]^2/2^3.r = [BC]^2/2^3.r$
ABC	1	$[(a-1)(b-1)(c-1)]^2/2^3.r = [ABC]^2/2^3.r$
Totals	7	S. S. (treatments)

(ii) By Yates's Method :

Treat. comb.	Total yield	I	II	III	T. F. effect	S. S.	due to
(1)	(1)	(1)+(b)	(1)+(a)+(b)+(ab)	(1)+(a)+(b)+(ab)+(c) +(ac)+(bc)+(abc)	G	$G^{2/2} \cdot r$ =C. F.	—
a	(a)	(b)+(ab)	(c)+(ac)+(bc)+(abc)	(a)—(1)+(ab)—(b)+(ac) —(c)+(abc)—(bc)	[A]	$[A]^{2/2} \cdot r$	A
b	(b)	(c)+(ac)	(a)—(1)+(ab)—(b)	(b)+(ab)—(1)—(a)+(bc) +(abc)—(c)—(ac)	[B]	$[B]^{2/2} \cdot r$	B
ab	(ab)	(bc)+(abc)	(ac)—(c)+(abc)—(bc)	(ab)—(b)—(a)+(1)+(abc) —(bc)—(ac)+(c)	[AB]	$[AB]^{2/2} \cdot r$	AB
c	(c)	(a)—(1)	(b)+(ab)—(1)(a)	(c)+(ac)+(bc)+(abc) —(1)—(a)—(b)—(ab)	[C]	$[C]^{2/2} \cdot r$	C
ac	(ac)	(ab)—(b)	(bc)+(abc)—(c)—(ac)	(ac)—(c)+(abc)—(bc) —(a)+(1)—(ab)+(b)	[AC]	$[AC]^{2/2} \cdot r$	AC
bc	(bc)	(ac)—(c)	(ab)—(b)—(a)+(1)	(bc)+(abc)—(c)—(ac) —(b)—(ab)+(1)+(a)	[BC]	$[BC]^{2/2} \cdot r$	BC
abc	(abc)	(abc)—(bc)	(abc)—(bc)—(ac)+(c)	(abc)—(bc)—(ac)+(c) —(ab)+(b)+(a)—(1)	[ABC]	$[ABC]^{2/2} \cdot r$	ABC

Exp. No. (2) The following lay out gives the barley yield (in kgms./plot) of 32 plots of a 2^3 factorial experiment in which three factors each at 2 levels are to be tested. Analyse the data and state your conclusions ? The notations are—

m for manure
 n for nitrogen and
 p for phosphorus

Replication	yield/plot (in kgms.)							
I	(1)	m	mn	n	p	np	mp	mnp
	30	45	50	24	27	34	40	74
II	m	(1)	p	n	mp	mnp	np	mn
	39	41	37	34	32	63	29	74
III	(1)	n	m	np	mp	mnp	mn	p
	25	22	46	30	38	68	54	25
IV	mn	m	p	n	(1)	mp	np	mnp
	62	43	26	16	17	46	28	61

Solution—

Ho : The data is homogeneous with respect to the treatments and the replications (blocks).

For convenience in calculations, we take the deviations from $y=40$ kgs. and prepare the following table to compute the $S. S.$ due to treatments and replications (blocks)—

Tr. comp. Rep.	I	m	n	mn	p	mp	np	mnp	Totals =B	(Totals) ² =B
I	-10 (100)	5 (25)	-16 (256)	10 (100)	-13 (169)	0 (0)	-6 (36)	34 (1156)	4 (1842)	16
II	1 (1)	-1 (1)	-6 (6)	34 (1156)	-3 (9)	-8 (64)	-11 (121)	23 (529)	23 (1917)	841
III	-15 (225)	6 (36)	-18 (324)	14 (196)	-15 (225)	-2 (4)	-10 (100)	26 (784)	-12 (1894)	144
IV	-23 (529)	9 (9)	-24 (576)	22 (484)	-14 (196)	6 (36)	-12 (144)	21 (441)	-21 (2415)	441
Totals =T	-47 (855)	13 (71)	-64 (1192)	80 (1936)	-45 (99)	-4 (104)	-39 (401)	106 (2910)	0=G (8068)	0
(Totals) ² =T	2209	169	4096	6400	2025	16	1521	11236	0	

$$C. F. = \frac{G^2}{N} = \frac{(0)^2}{32} = 0$$

$$T. S. S. = \sum_{ij} y_{ij}^2 - C. F. = 8068 - 0 = 8068$$

$$S. S. (\text{blocks}) = \sum_j \frac{B_j^2}{8} - C. F. = \frac{16+841+144+441}{8} - 0$$

$$= \frac{1142}{8} = 180.25$$

$$S. S. (\text{treat.}) = \sum_i \frac{T_i^2}{4} - C. F. = \frac{2209 + \dots + 11236}{4} - 0 = \frac{27672}{4}$$

$$= 6918.0$$

$$S. S. (\text{error}) = T. S. S. - S. S. \text{ due to (blocks + treat.)}$$

$$= 8068 - (180.25 + 6918.00)$$

$$= 8068 - 7098.25 = 969.75$$

The treatment-component sum of squares are obtained by Yate's Method—

Treat. comp.	Total yield	I	II	III	Total fact. effect	S. S.	due to
(1)							
<i>m</i>	-47 } 13 }	-34 } 16 }	-18 } 18 }	0 390	<i>G</i> [<i>M</i>]	$(0)^2/32 = 0 = C. F.$ $(390)^2/32 = 4753.125$	— <i>M</i>
<i>n</i>	-64 } 80 }	-49 } 67 }	204 } 186 }	166 188	[<i>N</i>] [<i>MN</i>]	$(166)^2/32 = 861.125$ $(188)^2/32 = 1104.500$	<i>N</i> <i>MN</i>
<i>p</i>	-45 } -4 }	60 } 144 }	50 } 116 }	36 -18	[<i>P</i>] [<i>MP</i>]	$(36)^2/32 = 40.500$ $(18)^2/32 = 10.125$	<i>P</i> <i>MP</i>
<i>np</i>	-39 } 106 }	41 } 145 }	84 } 10 }	66 20	[<i>NP</i>] [<i>MNP</i>]	$(66)^2/32 = 136.125$ $(20)^2/32 = 12.500$	<i>NP</i> <i>MNP</i>
						Total = 6918.00	

Now we arrive at the following A. V. T.—

Source of variation	D. F.	S. S.	M. S. S.	F cal.	F at	
					5%	1%
Blocks	3	180.250	60.0833	1.08	3.07	4.87
Treat.	7	6918.000	988.2857	21.41**	2.49	3.65
M	1	4753.125	4753.1250	102.902**	4.32	8.02
N	1	861.125	861.1250	18.65**	"	"
MN	1	1104.500	1104.5000	23.17**	"	"
P	1	40.500	40.5000	1.14	248.25	6214.5
MP	1	10.125	10.1250	4.56	"	"
NP	1	136.125	136.1250	2.95	4.32	8.02
MNP	1	12.500	12.5000	3.69	248.25	6214
Error	21	969.750	46.1798	—	—	—
Totals	31	8068.000	—	—	—	—

Inference : F-test indicates that the main effects M, N and the interaction MN are highly significant which leads to the conclusion that over all the application of manure and nitrogen increase the yields and they are not independent but interact each other.

Exp. (3) : In an N, P, K trial, with two levels of each fertilizer and 3 replicates, the treatment-totals were—

(1)	n	p	k	np	nk	pk	npk
94	108	97	98	114	123	111	124

The error mean square is known to be 8.90. Calculate the sum of squares for N and NP and test their significance ?

(M. Sc. Agrs, 1962)

Solution :

Ho : The effect of N and NP are not significant.

$$\begin{aligned}
 S. S. (N) &= \frac{[N]^2}{2^3 r} = \frac{[(n-1)(p+1) \cdot k + 1]^2}{8 \times 3} \\
 &= \frac{[(npk) + (nk) - (pk) - (k) + (np) + (n) - (p) - (1)]^2}{24} \\
 &= \frac{[124 + 123 - 111 - 98 + 114 + 108 - 97 - 94]^2}{24} \\
 &= \frac{(69)^2}{24} = 198.375
 \end{aligned}$$

$$\begin{aligned}
 S. S. (NP) &= \frac{[NP]^2}{2^3 r} = \frac{[(n-1)(p-1)(k+)]^2}{8 \times 3} \\
 &= \frac{[(npk) - (nk) - (pk) + (k) + (np) - (n) - (p) + (1)]^2}{24} \\
 &= \frac{[124 - 123 - 111 + 98 + 114 - 108 - 97 + 94]^2}{24} = \frac{(9)^2}{24} \\
 &= 3.375
 \end{aligned}$$

A. V. T.—

Source of variation	D. F.	S.	M. S. S.	F cal.	F at	
					5%	1%
N	1	198.375	198.375	22.29**	4.60	8.86
NP		3.375	3.375	2.63	245	6142
Error	14	—	8.90	—	—	—

Inference : The main effect of the fertilizer N is highly significant which proves that over all there is a significant response to N and the interaction NP is insignificant.

(iii) **Case of $m \times n$ factorial experiment :** A factorial experiment with 2 factors A and B at m and n levels respectively is called a $m \times n$ factorial experiment. Let the 'm' levels of A be denoted by $a_0, a_1, a_2, \dots, a_{m-1}$ and 'n' level B by $b_0, b_1, b_2, \dots, b_{n-1}$. There are 'mn' possible treatment combinations. If 'r' denotes the no. of replication then the symbol $(a_i b_j)$ denotes the yield of 'r' plots which receives the treatment combination $a_i b_j$.

The Partitioning of treatment S. S : The treatment *S. S.* with $(mn-1)$ d. f. will be split up into the following three components—

(1) *S. S.* due to *A* with $(m-1)$ d. f.,

(2) *S. S.* due to *B* with $(n-1)$ d. f.,

and (3) *S. S.* due to *AB* with $(m-1)(n-1)$ d. f. by arranging the total yields in the $m \times n$ table given below—

$\begin{array}{c} A \\ \diagdown \\ B \end{array}$	a_0	a_1	a_2	a_{m-1}	Totals
b_0	(a_0b_0)	(a_1b_0)	(a_2b_0)	$(a_{m-1}b_0)$	(b_0)
b_1	(a_0b_1)	(a_1b_1)	(a_2b_1)	$(a_{m-1}b_1)$	(b_1)
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
b_{n-1}	(a_0b_{n-1})	(a_1b_{n-1})	(a_2b_{n-1})	$(a_{m-1}b_{n-1})$	(b_{n-1})
Totals	(a_0)	(a_1)	(a_2)	(a_{m-1})	G

Now we compute

$$S. S. (\text{treat.}) = \sum_i^{m-1} \sum_j^{n-1} \frac{(a_{ij})^2}{r} - C. F.,$$

$$\text{where } C. F. = \frac{G^2}{N} \text{ and } N = r.m.n.$$

$$S. S. (A) = \sum_i \frac{(a_i)^2}{r.n} - C. F.$$

$$S. S. (B) = \sum_j \frac{(b_j)^2}{rm} - C. F.,$$

$$\text{and } S. S. (AB) = S. S. (\text{treat.}) - S. S. (A) - S. S. (B)$$

Standard Error : To compare the responses due to levels of A, B and different treatment combinations, we require three different s. E._s .—

(i) *S. E.* of difference between the two means of A

$$= \sqrt{\frac{2V_E}{nr}}$$

(ii) *S. E.* of the difference between the two means of B

$$= \sqrt{\frac{2V_E}{mr}}$$

and (iii) *S. E.* of the difference between the two means of treatment combinations

$$= \sqrt{\frac{2V_E}{r}}$$

Exp. No. (4) Three varieties of wheat (v_1, v_2 and v_3) and 4 doses of ammonium sulphate [none (n_0), 22 kgms./Hect. (n_1), 44 kgms./Hect. (n_2) and 66 kgms./Hect. (n_3)] were tested in a *R. B. D.* with 4 replicates. The layout with plot-yields (in kgms.) is given below. Analyse the data and state your conclusions ?

(plot size 1/80 acre)

Rep. I				Rep. II			
v_2n_2	v_3n_3	v_1n_0	v_3n_1	v_1n_2	v_1n_0	v_2n_1	v_1n_3
17	16	15	16	15	15	16	13
v_1n_3	v_3n_2	v_1n_2	v_2n_0	v_2n_3	v_3n_2	v_3n_0	v_2n_2
15	19	18	13	18	20	15	17
v_3n_0	v_2n_1	v_2n_3	v_1n_1	v_1n_2	v_3n_1	v_3n_3	v_2n_0
14	15	18	16	17	16	18	15
Rep. III				Rep. IV			
v_1n_2	v_2n_2	v_2n_1	v_2n_0	v_3n_0	v_1n_1	v_1n_2	v_3n_1
15	19	17	13	15	16	17	20
v_1n_1	v_1n_0	v_3n_3	v_1n_3	v_2n_3	v_2n_0	v_3n_2	v_1n_3
17	15	19	16	19	17	22	17
v_2n_3	v_3n_1	v_3n_2	v_3n_0	v_2n_2	v_3n_3	v_2n_1	v_1n_0
17	17	20	14	20	19	20	16

Solution—

Ho : The data is homogeneous with respect to the blocks (replicates) and the treatments.
Taking deviations from $y=17$, we Prepare the following table—

Treat. Rep.	v_{1no}	v_{2no}	v_{3no}	v_{1n_1}	v_{2n_1}	v_{3n_1}	v_{1n_2}	v_{2n_2}	v_{3n_2}	v_{1n_3}	v_{2n_3}	v_{3n_3}	Totals =B	Totals (Totals =B) ²
I	-2 (4)	-4 (16)	-3 (9)	-1 (1)	-2 (4)	-1 (1)	1 (1)	0 (0)	2 (4)	-2 (4)	1 (1)	-1 (1)	-12 (46)	144
II	-2 (4)	-2 (4)	-2 (4)	-2 (4)	-1 (1)	-1 (1)	0 (0)	0 (0)	3 (9)	-4 (16)	1 (1)	1 (1)	-9 (45)	81
III	-2 (4)	-4 (16)	-3 (9)	0 (0)	0 (0)	0 (0)	-2 (4)	2 (4)	3 (9)	-1 (1)	0 (0)	2 (4)	-5 (51)	25
IV	-1 (1)	0 (0)	-2 (4)	-1 (1)	3 (9)	3 (9)	0 (0)	3 (9)	5 (25)	0 (0)	2 (4)	2 (4)	14 (66)	196
Totals =T	-7 (13)	-10 (36)	-10 (26)	(6)	0 (14)	1 (11)	-1 (5)	5 (1)	13 (47)	-7 (21)	4 (6)	4 (10)	-12 (20)	14
(Totals =T) ²	49	100	100	16	0	1	1	25	169	49	16	16	144	

$$C. F. = \frac{G^2}{N} = \frac{144}{48} = 30$$

$$T. S. S. = \sum_{i,j} \sum y^2_{ij} - C. F. = 208 - 3 = 205.00$$

$$S. S. (treat.) = \sum_i \frac{T_i^2}{4} - C. F. = \frac{49+100+\dots+16}{4} - 3.0$$

$$= \frac{542}{4} - 3 = 135.5 - 3.0 = 132.50$$

$$S. S. (blocks) = \sum_j \frac{B_j^2}{12} - C. F. = \frac{144+81+25+196}{12} - 3.0$$

$$= \frac{546}{12} - 3.0 = 37.17 - 3.0 = 34.17$$

$$S. S. (error) = T. S. S. - S. S. (treat. + blocks)$$

$$= 205.00 - (132.50 + 34.17) = 205.00 - 166.67 = 38.33$$

Now the treatment *S. S.* can be partitioned into its components by forming the following table—

$\begin{array}{c} \diagup \\ N \\ \diagdown \\ V \end{array}$	n_0	n_1	n_2	n_3	Totals =V	$(\text{Totals})^2$ =V
v_1	-7	-4	-1	-7	-19	361
v_2	-10	0	5	4	-1	1
v_3	-10	1	13	4	8	64
Totals =N	-27	-3	17	1	-12	144
$(\text{Totals})^2$ =N	729	9	289	1	144	

$$S. S. (V) = \sum_i \frac{V_i^2}{16} - C. F. = \frac{361+1+64}{16} - 3.0 = \frac{426}{16} - 3.0$$

$$= 26.625 - 3.000$$

$$= 23.625$$

$$S. S. (N) = \sum_j \frac{N_j^2}{12} - C. F. = \frac{729+9+289+1}{12} - 3.0 = 85.667$$

$$S. S. (NV) = S. S. (\text{treat.}) - S. S. (V + N)$$

$$= 132.50 - (23.625 + 85.667) = 132.500 - 109.292$$

$$= 23.208$$

Finally we arrive at the following *A. V. T.*—

Source of variation	<i>D. F.</i>	<i>S. S.</i>	<i>M. S. S.</i>	<i>F cal.</i>	<i>F at</i>	
					5%	1%
Blocks	3	34.170	11.39	9.802**	2.89	4.44
Treat.	11	132.500	12.0455	10.36**	2.09	2.84
V	2	23.625	11.8125	10.16**	3.29	5.315
N	3	85.667	28.5557	24.57**	2.89	4.44
NV	6	23.208	3.868	3.32*	2.39	3.4
Error	33	38.33	1.162 = V_E	—	—	—
Totals	47	205.000	—	—	—	—

Inference—The calculated values of *F* corresponding to the blocks (replicates), treatment-components *N* and *V* come out to be highly significant showing that the main effects of ammonium sulphate and varieties of wheat are significant. The two factors (varieties of wheat and ammonium sulphate) are not independent to each other but they interact as the effect of interaction (*NV*) is also significant.

Now we compute the $S. E_s$ of the differences between the two means of V, N and NV by the following formulae—

(i) $S. E.$ of the difference between the two means of V.

$$= \sqrt{\frac{2V_E}{nr}} = \sqrt{\frac{2 \times 1.162}{4 \times 4}} = \sqrt{0.14525} = 0.3811$$

(ii) $S. E.$ of the difference between the two means of N.

$$= \sqrt{\frac{2V_E}{mr}} = \sqrt{\frac{2 \times 1.162}{3 \times 4}} = \sqrt{0.1937} = 0.4401$$

and (iii) $S. E.$ of the difference between the two means of NV.

$$= \sqrt{\frac{2V_E}{r}} = \sqrt{\frac{2 \times 1.162}{4}} = \sqrt{0.581} = 0.7622$$

The critical differences ($C. D_s$) for $S. E_s$ given in (i), (ii) &

(iii) will be computed in the following manner—

(i) ($C. D.$) = ($S. E.$ of difference) $\times t$ (33)

$$\begin{array}{cc} 5\% & .05 \\ = 0.3811 \times 3.29 = 1.2538 \text{ kgms./plot} \end{array}$$

(ii) ($C. D.$) = ($S. E.$ of difference) $\times t$ (33)

$$\begin{array}{cc} 5\% & .05 \\ = 0.4401 \times 3.29 = 1.4479 \text{ kms./plot} \end{array}$$

and (iii) ($C. D.$) = ($S. E.$ of difference) $\times t$ (33)

$$\begin{array}{cc} 5\% & .05 \\ = 0.7622 \times 3.29 = 2.5076 \text{ kgms./plot} \end{array}$$

Effect of ammonium sulphate—

doze of N (in kgms./Hect.)	average yield/plot (in kgms.)	average yield/acre (in kgms.)
0	14.75	$14.75 \times 90 = 1180$
22	16.75	$16.75 \times 80 = 1340$
44	18.4167	$18.4167 \times 80 = 1473.336$
66	17.0833	$17.0833 \times 80 = 1366.664$

$$\begin{aligned} C. D. \text{ at } 5\% \text{ for ammonium sulphate-means} &= 1.4479 \times 80 \\ &= 115.832 \text{ kgms./acre} \end{aligned}$$

An inspection of the above data reveals the fact that the application of ammonium sulphate has exhibited a significant effect on the yields. The yields obtained by applying ammonium sulphate at the rate of 22 kgms./Hect., 44 kgms./Hect., and 66 kgms./Hect. are higher than that obtained without applying it. The rate of 44 gms./hect. is the best of all.

Varietal effect—

Variety of wheat	average yield per plot (in kgms/Hect)	average yield per acre (in kgms/acre)
v_1	12.25	$12.25 \times 80 = 980$
v_2	16.9167	$16.9167 \times 80 = 1353.336$
v_3	17.67	$17.67 \times 80 = 1413.6$

C. D. at 5% for variety of wheat $= 1.2538 \times 80 = 100.304$

kgms/acre

The main effect of variety (V) is highly significant which proves that over all there is a significant response to V . The maximum yield has been recorded in the case of variety v_3 followed by v_2 but their difference is not significant. The varieties v_2 and v_3 are significantly different from v_1 .

The effect of NV—

(average yields in kgms/acre)

N V	n_0	n_1	n_2	n_3
v_1	$(-\frac{7}{4} + 17) \times 80 = 1220$	1280	1340	1220
v_2	1160	1360	1460	1440
v_3	1160	1380	1620	1440

The C. D. at 5% for interaction (NV) $= 2.5076 \times 80$

$= 200.608$ kgms/acre

The interaction (NV) is significant at 5% level which shows that over all the treatment combinations differ significantly. The maximum yield has been recorded for the treatment combination n_2v_3 followed by n_2v_2 but their difference is not significant. The variety v_3 with ammonium sulphate at the rate of 44 kgmS/Hect. is the best combination.

Solution—

Ho : The main effects (M and V) and interaction (MV) are not significant.

Exp. No. (5) In a manurial cum variety trial (3 manures \times 2 varieties) laid out in a randomized block design with four replications, the following treatment totals are obtained—

Totals of 4 plots

Treatments : v_1m_1 v_1m_2 v_1m_3 v_2m_1 v_2m_2 v_2m_3

Totals : 21 30 40 27 40 42

The residual variance for 15 D. $F = 1.5450$. Calculate the main effects (M and V) and interaction (MV) and test their significance ?

(M. Sc. Ag. Agra, 1959)

To compute the S. S. for treatments, main effects and interaction, we perform the following table—

$\begin{array}{c} \text{M} \\ \diagdown \\ \text{V} \end{array}$	m_1	m_2	m_3	Totals =M	(Totals =M) ²
v_1	21 (441)	30 (900)	40 (1600)	91 (2941)	8281
v_2	27 (729)	40 (1600)	42 (1764)	109 (4093)	11881
Totals=V	48 (1170)*	70 (2500)	82 (3364)	200=G (7034)	40000
$\left(\begin{array}{c} \text{Totals} \\ =V \end{array}\right)^2$	2304	4900	6724	40000	

$$C. F. = \frac{G^2}{N} = \frac{40000}{24} = 1666.67$$

$$\begin{aligned} \text{since } N &= r \times m \times n \\ &= 4 \times 3 \times 2 \\ &= 24 \end{aligned}$$

$$\begin{aligned} S. S. (\text{treat.}) &= \sum_i \sum_j \frac{y_{ij}^2}{4} - C. F., \text{ where } y_{ij} \text{ is the yield of 4 plots.} \\ &= \frac{7034}{4} - 1666.67 = 1758.50 - 1666.67 = 91.83 \end{aligned}$$

$$\begin{aligned} S. S. (M) &= \sum_i \frac{M_i^2}{4} - C. F. = \frac{2304 + 4900 + 6724}{8} - 1666.67 \\ &= \frac{13928}{8} - 1666.67 \\ &= 1741 - 1666.67 \\ &= 74.33 \end{aligned}$$

$$\begin{aligned} S. S. (V) &= \sum_j \frac{V_j^2}{12} - C. F. = \frac{8281 + 11881}{12} - 1666.67 \\ &= \frac{20162}{12} - 1666.67 = 1680.167 - 1666.67 \\ &= 13.497 \end{aligned}$$

$$\begin{aligned} S. S. (MV) &= S. S. (\text{treat.}) - S. S. (M + V) \\ &= 91.83 - (74.33 + 13.497) = 91.83 - 87.827 \\ &= 4.003 \end{aligned}$$

Now we prepare the following A. V. T.—

Source of variation	D. F.	S. S.	M S. S.	F cal.	F at	
					5%	1%
Treatment	5	91.83	18.366	11.82**	2.90	4.56
M	2	74.33	37.165	24.05**	3.68	6.36
V	1 = 5	13.497	13.497	8.74**	4.54	8.68
MV	2	4.003	2.0015	1.29	3.68	6.36
Error	15	—	1.545	—	—	—

Inference—The calculated variance ratios for treatment main-effects M and V come out to be highly significant which show that the main-effects of manures and varieties are significant. The two factors (manure and variety) are independent of each other and do not interact as their interaction effect is not significant.

Exp. No. (6) An experiment was planned to study the effect of ammonium sulphate (N) and super phosphate (P) on the yield of maize (Ganga hybrid No. 1). All combinations of four levels of super phosphate (0 kgms, 10 kgms, 20 kgms and 30 kgms/acre) and ammonium sulphate (0 kgms, 10 kgms and 20 kgms/acre) were studied in a R. B. D. with 2 replications. The following yields (kgms/plot, plot size=1/40 acre) were obtained—

Treat.	Rep. I	Rep. II	Totals
n_0P_0	181.2	184.2	365.4
n_0P_1	184.5	184.5	369.0
n_0P_2	191.5	191.9	383.4
n_0P_3	199.8	203.2	403.0
n_1P_0	196.9	211.3	408.2
n_1P_1	192.5	212.6	405.1
n_1P_2	188.9	201.3	390.2
n_1P_3	200.5	207.4	407.9
n_2P_0	215.2	231.8	447.0
n_2P_1	208.2	216.2	424.4
n_2P_2	109.8	208.5	418.3
n_2P_3	221.8	221.5	443.3

Obtain the M. S. for N, P and NP when it is given that the error variance in this experiment is 27.363. Test the significance of N, P and NP ?

Solution—

Ho : The effects of N, P and NP are not significant.

In order to obtain the S. S. for N, P and NP, we arrange the data in the following 4×3 table after taking the deviations from $y=400$ kgrms.

$\begin{array}{c} F \\ \diagdown \\ N \end{array}$	P_0	P_1	P_2	P_3	Totals = P	(Totals = P) ²
N_0	$\begin{array}{c} -34.6 \\ (1197.16) \end{array}$	$\begin{array}{c} -31.0 \\ (961) \end{array}$	$\begin{array}{c} -16.6 \\ (275.56) \end{array}$	$\begin{array}{c} 3.0 \\ (9) \end{array}$	$\begin{array}{c} -79.2 \\ (2442.72) \end{array}$	6272.64
N_1	$\begin{array}{c} 8.2 \\ (67.24) \end{array}$	$\begin{array}{c} 5.1 \\ (26.01) \end{array}$	$\begin{array}{c} -9.8 \\ (96.04) \end{array}$	$\begin{array}{c} 7.9 \\ (62.41) \end{array}$	$\begin{array}{c} 11.4 \\ (251.70) \end{array}$	129.96
N_2	$\begin{array}{c} 47.0 \\ (2209.00) \end{array}$	$\begin{array}{c} 24.4 \\ (595.36) \end{array}$	$\begin{array}{c} 18.3 \\ (334.89) \end{array}$	$\begin{array}{c} 43.3 \\ (1874.89) \end{array}$	$\begin{array}{c} 133.0 \\ (5014.14) \end{array}$	17689
Totals = N	$\begin{array}{c} 20.6 \\ (3473.40) \end{array}$	$\begin{array}{c} -1.5 \\ (1582.39) \end{array}$	$\begin{array}{c} -8.1 \\ (706.49) \end{array}$	$\begin{array}{c} 54.2 \\ (1946.30) \end{array}$	$\begin{array}{c} 65.2=G \\ (7708.58) \end{array}$	4251.04
(Totals = N) ²	424.36	2.25	65.61	2937.64	4251.04	

$$C. F. = \frac{G^2}{N} = \frac{4251.04}{24} = 177.1267$$

$$S. S. (treat.) = \sum_i \sum_j \frac{y_{ij}^2}{2} - C. F., \text{ where } y_{ij} \text{ are the yields of } 2 \text{ plots}$$

$$= \frac{7708.56}{2} - 177.1267 = 3854.28 - 177.1267$$

$$= 3677.1533$$

$$S. S. (P) = \sum_i \frac{P_i^2}{6} - C. F. = \frac{424.36 + \dots + 2937.64}{6} - 177.1267$$

$$= \frac{3429.83}{6} - 177.1267$$

$$= 571.6383 - 177.1267$$

$$= 394.5116$$

$$S. S. (N) = \sum_j \frac{N_j^2}{8} - C. F. = \frac{6272.64 + 129.96 + 17689.00}{8} - 177.1267$$

$$= \frac{24091.60}{8} - 177.1267$$

$$= 3011.45 - 177.1267$$

$$= 2834.3233$$

$$S. S. (NP) = S. S. (treat) - S. S. \text{ due to } (N+P)$$

$$= 3677.1533 - (2834.3233 + 394.5116)$$

$$= 3677.1533 - 3228.8349 = 448.3184$$

Now we arrive at the following A. V. T.—

Source of variation	D. F.	S. S.	M. S. S	F cal.	F at	
					5%	1%
P	3	394.5116	131.5039	4.81*	3.59	6.22
N	2	2834.3233	1417.1617	51.79**	3.98	7.00
NP	6	448.3184	74.7197	2.73	3.09	5.07
Error	11	—	27.363 = V_E			

Inference—*F*-test indicates that the main effect *P* is significant at 5% and *N* is highly significant while the interaction *NP* is not significant.

For comparing the response exhibited by different levels of *P* and *N*, we compute the *C. D.*_s as given below—

(i) *S. E.* of the difference between the two means of *P*

$$= \sqrt{\frac{2V_E}{nr}} = \sqrt{\frac{2 \times 27.363}{3 \times 2}} = \sqrt{9.121} = 3.02$$

$$(C. D.)_{5\%} = (S. E. \text{ of difference}) \times t_{.05} \quad (11)$$

$$= 3.02 \times 2.201 = 6.647 \text{ kgms/plot}$$

(ii) *S. E.* of the difference between the two means of *N*.

$$= \sqrt{\frac{2V_E}{mr}} = \sqrt{\frac{2 \times 27.363}{4 \times 2}} = \sqrt{6.84075} = 2.6154$$

$$(C. D.)_{5\%} = (S. E. \text{ of difference}) \times t_{.05} \quad (11)$$

$$= 2.6154 \times 2.201 = 5.7565 \text{ kgms/plot}$$

Effect of super phosphate—

doze of <i>P</i> (in kgms/acre)	average yield/plot (in kgms.)	average yield/acre (in kgms.)
0	203.43	$203.43 \times 40 = 8137.20$
10	199.75	$199.75 \times 40 = 7990.0$
20	198.65	$198.65 \times 40 = 7946.0$
30	209.03	$209.03 \times 40 = 8361.20$

$$\begin{aligned} C. D. \text{ at } 5\% \text{ for super phosphate means} &= 6.647 \times 40 \\ &= 265.88 \text{ kgms/acre} \end{aligned}$$

The application of super-phosphate at the rates of 10, 20 and 30 kgms/acre do not exhibit significant effects as compared to control while the effect of applying 30 kgms/acre is significantly different from the effects of the doses 10 and 20 kgms/acre. However, the max. yield has been recorded in the case of 30 kgms/acre.

Effect of ammonium Sulphate-

doze of N (in kgms/acre)	average yield/plot (in kgms.)	average yield/acre (in kgms.)
3	190.1	$190.1 \times 40 = 7604.0$
10	201.425	$201.425 \times 40 = 8057.0$
20	216.625	$216.625 \times 40 = 8665.0$

$$\begin{aligned}
 C. D. \text{ at } 5\% \text{ for ammonium sulphate means} &= 2.6154 \times 40 \\
 &= 230.26 \text{ kgms/acre}
 \end{aligned}$$

The application of ammonium-sulphate at the rates of 10 and 20 kgmn/acre differ significantly from that of control. The maximum yield has been recorded in the case of 20 gms/acre followed by 10 kgms/acre and their difference is significant.

Factorial approach Versus Traditional approach.

S.N.	Factorial Approach	Traditional Approach
1	It can be used in either case, whether the factors are independent or intetract each other. Thus, it can be applied to a wide variety of cases.	It can be used only when the factors are independent. Thus, its use is limited.
2	It compares all the factors in one and the same experiment and thus saves a great deal of time and the experimental material.	It compares different factors in the different, single, independent experiments and thus requires more time and the experimental material.
3	It provides the informations regarding the main-effects as well as interactions.	It gives the information about the main-effects only.
4	It compares all the factors with the same precision and randers the comparison of the results.	It compares the different factors with different precisions and does not render tho comparison of the results.
5	It selects the optimum combination.	It cannot select the optimum combination.
6	It compares each factor under the varying conditions of the other factors and hence the scope of the experiment is greater.	It compares each factor by keeping other factors constant and hence the scope of the experiment is limited.

Thus, it is evident that the factorial experiments are of greater efficiency and comprehensiveness.

EXERCISE VII

Q. 1. Describe the factorial method of experimentation and explain advantages ?

What factors do you consider worth trial when testing four new cotton varieties ? Describe briefly the design you would choose for the trial and the layout of the experiment ?

(M. Sc. Ag. Agra, 1958)

Q. 2. An experiment is to be conducted for finding the effect of the three levels of irrigation on the yield of four varieties of potatoes. Suggest a suitable plan for this experiment and give the skelton of analysis of variance table ? Also indicate the method for calculating the sums of squares for the different components ?

(M. Sc. Ag. Agra, 1961)

Q. 3. The following table gives the treatment totals for an experiment with three levels of nitrogen fertilizer and three levels of phosphate fertilizer. The data are the number of lettuce plants that emerged from the ground and are totals over 12 plots—

Number of lettuce plants emerging—

		Levels of nitrogen			Totals
		n_0	n_1	n_2	
Levels of phosphate	p_0	449	413	326	1188
	p_1	409	358	291	1058
	p_2	341	278	312	931
Totals		1199	1049	929	3177

(a) Obtain the m. s. for nitrogen, phosphate and nitrogen \times phosphate ?

(b) It is given that the error m.s. in this experiment is about 59. Which of the effects indicate significance ?

Ans. (a) m. s. for

$N=508.3334$, $P=399.8056$

and $NP=79.2639$

(b) N and P are significant.

Q 4. An experiment was planned to study the effect of sulphate of potash and super phosphate on yield of potatoes. All combinations of the three levels of super sulphate [0 cwt (p_0), 5 cwt (p_1), 10 cwt (p_2) per acre] and two levels of sulphate of potash [0 cwt (k_0), 2 cwt (k_1) per acre] were studied in a 6×6 L. S. The following yields (lbs/plot = 1/70 acre) were obtained—

6 × 6 Latin Square

k_1p_0 186	k_0p_2 187	k_0p_0 208	k_0p_1 222	k_1p_1 296	k_1p_2 331
k_1p_1 213	k_0p_0 134	k_1p_2 296	k_0p_2 265	k_0p_1 250	k_1p_0 253
k_0p_1 198	k_1p_0 155	k_0p_2 272	k_1p_2 290	k_0p_0 261	k_1p_1 310
k_1p_2 233	k_1p_1 184	k_0p_1 218	k_1p_0 234	k_0p_2 248	k_0p_0 293
k_0p_2 245	k_1p_2 233	k_1p_1 282	k_0p_0 248	k_1p_0 247	k_0p_1 303
k_0p_0 196	k_0p_1 228	k_1p_0 242	k_1p_1 255	k_1p_2 273	k_0p_2 294

Analyse the data and give your conclusions ?

(M. A. Patna, 1954)

Ans. $V_E = 340.7835$ for 20 d. f.

$$V_R = 820.25$$

$$V_C = 7876.45$$

$$V_T = 1363.135$$

$$V_K = 2288.03$$

$$V_P = 6996.335$$

$$V_{KP} = 265.44$$

CHAPTER VIII

Confounding

In factorial experiments, when the no. of factors and their levels is large i. e. the treatment combinations are numerous, the blocks require a larger area in comparison to that required for a fewer no. of factors and their levels. It is a common experience in agricultural experimentation that within a large area considerable soil heterogeneity is present which increases the experimental error and lowers the precision of the experiment. Thus, when the no. of treatments is numerous, the precision of the factorial experiment is affected adversely. One method of re-introducing the homogeneity within the blocks and of increasing the precision is to partition each replicate into two or more blocks such that the main effects and the interactions of interest are tested with a relatively higher precision than the interactions of little experimental value. This object is achieved by adopting the *artifice of confounding*. *It consists in mixing up inseparably the effects of unimportant interactions with the block effects.*

Technique of Confounding in 2^3 factorial experiment : In fact, the confounding is not necessary in this case but it has been chosen for the convenience of ready grasp of the technique. Suppose, the three factors are A, B and C each at 2 levels. The possible treatment combinations are (1), a, b, ab, c, ac, bc and abc. In order to

confound abc (three factor interaction), first we write down its corresponding algebraic expression $(a-1)(b-1)(c-1) = abc - bc - ac + a - ab + b + c - (1)$ and then randomize the four treatment-combinations of +ve sign (abc , b , c and a) in one block of each replicate and the remaining four of -ve sign (bc , ac , ab , 1) in the other block. Each replicate consists of 2 blocks. In a similar manner, any main effect or interaction can be confounded with the block effects. The effect of the confounded interaction is inseparably mixed with the blocks effect and hence it cannot be estimated separately. The unconfounded main effects and interactions are independent of the blocks-effect and can be estimated separately.

In the confounding scheme, the precisions of the unconfounded main effects and the interactions are higher than the precisions of the confounded ones. Hence, relatively unimportant interactions should be confounded. Generally, the higher order interactions are deemed to be unimportant either because of their insignificance or because of the impossibility of the application of several factors-treatment combinations.

Complete confounding : If the same interaction has been confounded with all the replications then it is a case of *complete or total confounding* and the interaction is called to be completely confounded. The precision regarding the completely confounded interaction is sacrificed altogether while that of regarding the others is increased. The complete confounding is chosen only when one of the interactions is of little importance and we do not require any knowledge regarding it from the experiment.

Statistical Analysis of completely confounded 2^3 factorial Expt. : The statistical analysis is carried out in the same way as in the case of factorial experiments using *yates method* for computing the $S.S.$ for main effects and interactions. The only modification is that neither the $S.S.$ for the completely confounded interaction is

computed nor it is included in the A. V. T. The skeleton of the A. V. T. when 'abc' is completely confounded, is given below—

Source of variation	D. F.	S. S.	M. S. S.	F cal.	F at	
					5%	1%
Blocks	$2r-1$
Treatments	6
<i>A</i>	1
<i>B</i>	1
<i>AB</i>	1
<i>C</i>	1
<i>AC</i>	1
<i>BC</i>	1
Error	$6(r-1)$
Totals	$8r-$	---	—	---	---

Exp. (1): A manurial trial was carried out on potato with N, P and K each at 2 levels. The interaction NPK is completely confounded with blocks. The plan and yields (in seers) are given below—

Rep. I		Rep. II		Rep. III	
B_1	B_2	B_3	B_4	B_5	B_6
<i>npk</i>	<i>np</i>	<i>k</i>	<i>kp</i>	<i>p</i>	<i>nk</i>
50	63	55	49	63	57
<i>p</i>	(1)	<i>n</i>	<i>np</i>	<i>k</i>	(1)
45	47	60	52	55	55
<i>n</i>	<i>nk</i>	<i>npk</i>	(1)	<i>npk</i>	<i>np</i>
62	57	60	51	55	60
<i>k</i>	<i>kp</i>	<i>p</i>	<i>nk</i>	<i>n</i>	<i>kp</i>
45	50	56	50	70	60

Analyse the data and state your conclusions ?

Solution :

Ho : The data is homogeneous with respect to blocks and treatments.

Taking the deviations from $y=60$ seers, we prepare the following tables to compute the S. S. for blocks, treatments and error—

	Rep. I	Rep. II	Rep. III	
Blocks treat.	B_1	B_3	B_5	Totals = T
npk	-10 (100)	0 (0)	-5 (25)	-15 (125)
p	-15 (225)	-4 (16)	3 (9)	-16 (250)
n	2 (4)	0 (0)	10 (100)	12 (104)
k	-15 (225)	-5 (25)	-5 (25)	-25 (275)
Totals = B	-18 (554)	-9 (41)	3 (159)	-44 (754)
(Totals = B) ²	1444	81	9	—
np	B_2 3 (9)	B_4 -8 (64)	B_6 0 (0)	-5 (75)
(l)	-13 (169)	-9 (81)	-5 (25)	-27 (275)
nk	-3 (9)	-10 (100)	-3 (9)	-16 (118)
kp	-10 (100)	-11 (121)	0 (0)	-21 (221)
Totals = B	-23 (287)	-38 (366)	-8 (34)	-69 (687)
(Totals = B) ²	529	1444	64	—

$$C.F. = \frac{G^2}{N} = \frac{[-44 + (-69)]^2}{24} = \frac{(113)^2}{24} = 532.0417$$

$$T.S.S. = \sum \sum_{ij} y_{ij}^2 - C.F. = (754 + 687) - 532.0417$$

$$= 908.9583$$

$$S.S. (\text{blocks}) = \sum_j \frac{B_j^2}{4} - C.F.$$

$$= \frac{(1444 + 81 + 9) + (529 + 1444 + 64)}{4} - 532.0417$$

$$= \frac{3571}{4} - 532.0417$$

$$= 892.75 - 532.0417 = 360.7083$$

Treatment S. S.—

Treat. comb.	Total yield	I	II	III	Total factorial effect	S. S.	due to
(1)	-27	-15	-36	-113	G	$(-113)^2/24$ $= 532.0417$	—
n	12	-21	-77	65	[N]	$(65)^2/24$ $= 176.0417$	N
p	-16	-41	50	-1	[P]	$(-1)^2/24$ $= 0.0417$	P
np	-5	-36	15	-31	[NP]	$(-31)^2/24$ $= 40.0417$	NP
k	-25	39	-6	-41	[K]	$(-41)^2/24$ $= 70.0417$	K
nk	-16	11	5	-35	[NK]	$(-35)^2/24$ $= 51.0417$	NK
pk	-21	9	-28	11	[PK]	$(11)^2/24$ $= 5.0417$	PK
npk	-15	6	-3	25	[NPK]	Confounded	NPK

$$S.S. (\text{treat.}) = 176.0417 + \dots + 5.0417 = 342.2502$$

$$S.S. (\text{error}) = T.S.S. - S.S. \text{ due to (blocks + treat.)}$$

$$= 908.9583 - (360.7083 + 342.2502)$$

$$= 908.9583 - 702.9585 = 205.9998$$

Finally we arrive at the following *A. V. T.*—

Source of variation	D. F.	S. S.	M. S. S.	F. cal.	F at	
					5%	1%
Blocks	5	360.7083	72.1417	4.20*	3.11	5.06
Treat.	6	342.2502	57.0417	3.32*	3.00	4.82
N	1	176.0417	176.0417	10.25**	4.75	9.33
P	1	0.0417	0.0417	411.75*	244.0	6106.0
NP	1 = 6	40.0417	40.0417	2.33	4.75	9.33
K	1	70.0417	70.0417	4.08	„	„
NK	1	51.0417	51.0417	2.97	„	„
PK	1	5.0417	5.0417	34.06	244.0	6106.0
Error	12	205.9998	17.17	—	—	—
Totals	23	908.9583	—	—	—	—

Inference—The blocks and the treatment effects are significant at 5% level and the treatment component N is highly significant while the other treatment components come out to be insignificant except P which is significant at 5% level only.

Partial Confounding—The confounding scheme in which the different interactions are confounded in different replicates is called *partial confounding*. Below is given an example of partial confounding where none of the interactions is cofounded in all the replicates—

Rep. I		Rep. II		Rep. III	
B ₁	B ₂	B ₃	B ₄	B ₅	B ₆
c	ab	abc	ac	a	ab
a	ac	c	bc	abc	ac
abc	(1)	ab	a	bc	b
b	bc	(1)	b	(1)	c
ABC confounded		AB confounded		BC confounded	

The interactions ABC , AB and BC are said to be partially confounded as any one of them is confounded in one replicate only. In partial confounding scheme any interaction is confounded in one or more replicates but not in all.

The above confounding scheme provides the full information regarding the unconfounded effects (A , B , C and AC) and partial ($2/3$) information regarding the confounded interactions (ABC , AB and BC). Because the unconfounded effects are estimated from all the three replicates while each of the confounded interactions is estimated from the 2 replicates only in which it is not confounded. Hence, the confounded interactions utilize $2/3$ observations while the unconfounded effects utilize the whole data.

Partial confounding is adopted when we want to increase the precision by partitioning the replicates into two or more blocks and at the same time desire to obtain the informations about all the effects but ready to sacrifice a fraction of informations regarding some interactions. The effects of greater importance are kept unconfounded while the effects of relatively less importance are confounded partially.

In the case of partial confounding in $2^3 F$. Expt., the treatments sum of squares are obtained from Yates Method and finally adjusting for the confounded effects. The adjusting factor for any confounded interaction is computed as below—

(i) Note the replicates in which the given interaction is confounded,

(ii) Note the sign of (1) in the corresponding algebraic expression. If the sign is +ve, then—

adjusting factor = [total of the blocks containing (1) of the replicates in which the interaction is confounded]—[total of the blocks not containing (1) of the replicates in which the interaction is confounded]

and if the sign of (1) is -ve, then

adjusting factor = $\frac{\text{[total of the blocks not containing (1) of the replicates in which the interaction is confounded]}}{\text{[total of the blocks containing (1) of the replicates in which the interaction is confounded]}}$

(iii) Keep the divisor $2^3 (r-i)$ where i is the no. of replicates in which the interaction is confounded.

The whole procedure of statistical analysis in a 2^3 partial confounding experiment is illustrated in the following example—

Exp. (2): The table given below gives the yields of wheat (in seers) per plot for a partially confounded 2^3 factorial experiment on 24 plots—

Rep. I				Rep. II		Rep. III	
B_1	B_2	B_3	B_4	B_5	B_6		
ab 101	b 88	(1) 125	ab 115	bc 75	a 53		
abc 111	a 90	abc 95	c 95	ac 100	abc 76		
(1) 75	bc 115	ac 80	bc 90	(1) 55	b 65		
c 55	ac 75	b 100	a 80	ab 95	c 82		

AB confounded

AC confounded

ABC confounded

Analyse the data and state your conclusions ?

Sol. :

• **Ho :** The data is homogeneous with respect to blocks and treatments.

Taking deviations from $y=87$ seers, we prepare the following table to compute the *S. S.* for blocks and *T. S. S.*—

Block. Treat. Comb.	Rep. I		Rep II		Rep III		Totals= <i>T</i>
	<i>B</i> ₁	<i>B</i> ₂	<i>B</i> ₃	<i>B</i> ₄	<i>B</i> ₅	<i>B</i> ₆	
(1)	-12 (144)	---	38 (1444)	---	-32 (1024)	---	-6 (2612)
a	---	3 (9)	---	-7 (49)	---	-34 (1156)	-38 (1214)
b	---	1 (1)	13 (169)	---	---	-22 (484)	-8 (654)
ab	14 (196)	---	---	28 (784)	5 (25)	---	47 (1005)
c	-32 (1024)	---	---	8 (64)	---	-5 (25)	-29 (1113)
ac	---	-12 (144)	-7 (49)	---	13 (169)	---	-6 (362)
bc	---	28 (784)	---	3 (9)	-12 (144)	---	19 (93)
a c	24 (576)	---	8 (64)	---	---	-11 (121)	21 (761)
Totals= <i>B</i>	-6 (1940)	20 (938)	52 (1726)	32 (906)	-26 (1362)	-72 (1786)	0= <i>G</i> (8658)
(Total= <i>B</i>) ²	36	400	2704	1024	676	5184	

$$C. F. = \frac{G^2}{N} = \frac{(0)^2}{24} = 0$$

$$T. S. S. = \sum_{ij} y_{ij}^2 - C. F. = 8658 - 0 = 8658$$

$$S. S. (\text{blocks}) = \sum_j \frac{B_j^2}{4} - C. F. = \frac{10024}{4} - 0 = 2506$$

For treatment S. S.

Treat. Comb.	Total yield	I	II	III	adjusting factor	adj. total fact. effect	S. S.	due to
(1)	-6	-44	-5	0	---	0	$C. F. = 0$	—
a	-38	39	5	48	---	48	96.00	A
b	-8	-35	23	158	---	158	1040.1667	B
ab	47	40	25	66	$B_1 - B_2 = -26$	$66 - (-26) = 92$	529.00	AB
c	-29	-32	83	10	---	10	4.4167	C
ac	-6	55	75	2	$B_3 - B_4 = 20$	$2 - 20 = -18$	20.25	AC
bc	19	23	87	-8	---	-8	2.6667	BC
abc	21	2	-21	-108	$B_5 - B_6 = -46$	$-108 - (-46) = -62$	240.25	ABC

$$S. S. (\text{total}) = 1932.750$$

$$S. S. (\text{error}) = T. S. S. - S. S. \text{ due to (blocks + treat.)}$$

$$= 8658 - (2506 + 1932.7501) = 8658 - 4438.7501$$

$$= 4219.2499$$

Finally we prepare the A. V. T.—

Source of variation	D. F.	'S.'S.	M. S. S	F. cal.	F at	
					5%	1%
Blocks	5	2506.00	501.2	1.31	3.20	5.32
Treat.	7	1932.75.1	276.1071	1.39	3.60	6.54
A	1	96.00	96.00	3.995	243	6082
B	1	1040.1667	1040.1667	2.71	4.84	9.65
AB	1	529.00	529.00	1.34	"	"
C	1	4.4167	4.4167	86.844	243	6082
AC	1	20.2500	20.25	18.94	"	"
BC	1	2.6667	2.6667	143.84	"	"
ABC	1	240.2500	240.25	1.60	"	"
Error	11	4219.2499	383.5682	—	—	—
Totals	23	8658.0	—	—	—	—

Inference : The data is homogeneous.

S. E. : The standard error of a mean for unconfounded effect

$$= \sqrt{\frac{V_E}{2r}} \text{ and that of a mean for confounded interaction}$$

$$= \sqrt{\frac{V_E}{2(r-i)}}$$

Where 'i' is the no. of replicates in which the interaction is confounded.

Merits and Demerits : The only advantage of confounding scheme lies in the fact that it reduces the experiment error considerably by partitioning the whole replicate into too or more blocks.

The dis-advantages are as follows—

(i) The confounded interactions can be estimated with lower precisions as the no. of replication for them is reduced.

(ii) The statistical analysis is complex and especially when some of the units (observations) are missing.

In confounding scheme, the increased precision is obtained at the cost of sacrifice of information (partially or completely) on certain unimportant interactions *which leads to the universal truth—that nothing can be achieved without the sacrifice of the other and certain relatively unimportant things should be sacrificed.*

Exercise VIII

Q. (1) : What is confounding ? An experiment is to be carried out with fertilizer N, P and K, each at two levels, on a crop, the triple interaction between the fertilizer being of no interest. Plan the layout of experiment in a confounded design and give a skelton analysis of the results. Will you recommend a confounded design for such a trial ? Give reasons ?

(M. Sc. Ag. Agra, 1956)

Q. (2) : What is confounding and what are its advantages ? A manurial trial is to be laid out in potato with N, P and K each at two levels. The interaction PK is to be confounded with blocks. Give the complete plan for this experiment, taking three replications. Give also the skelton analysis of variance ?

(M. Sc. Ag. Agra, 1959)

Q. (3) : A factorial experiment involving 3 fertilizers N, P & K each at 2 levels was carried out in 6 replicates and the second order interaction NPK is completely confounded. The following results were recorded—

Treat.	:	(1)	n	p	np	k	nk	pk	npk
Total yield :									
of 6 plots		971	1106	1219	1045	917	1187	1172	1203
(in kgms)									=8820

Calculate the m. s. for all the unconfounded effects and test their significance ? When the error m. s. is 1175.3304

Ans. Effect : N P NP K NK PK
M. S. : 1430.0833 4370.0833 5056.3333 396.75 2408.3383 147.00

Q. (4) : Analyse the following 2^3 partial confounding experiment—

Rep. I		Rep. II		Rep. III	
B_1	B_2	B_3	B_4	B_5	B_6
abc	b	a	bc	ab	(1)
114	62	93	90	72	75
ab	bc	b	ab	bc	abc
92	82	84	90	83	60
(1)	ac	c	ac	c	ac
83	92	87	75	83	75
c	a	abc	(1)	a	b
73	52	71	88	100	92

AB confounded ABC confounded AC confounded

Ans.

Source : Block, A, B, AB, C, AC, BC, ABC Error
m. s. 19.03, 0.67, 10.67, 68.06, 0.17, 6.25 8.17, 20.25, 284.75

Q. (5) - Analyse the following 2^3 partial confounding experiment—

Rep. I	
B_1	B_2
(1)	n
30	30
p	np
15	12

N-confounded

Rep. II	
B_3	B_4
(1)	p
30	10
n	np
26	10

P-confounded

Rep. III	
B_5	B_6
(1)	n
25	40
np	p
16	20

NP-confounded

Ans.

Source :	Blocks,	N,	P,	NP,	Error
M. S. :	99.0,	6.125,	480.50,	0.125,	12.0833

CHAPTER IX

Split Plot Design

Let us suppose that there are two factors A and B with m and n levels respectively and the factor A cannot be tested on small amount of experimental material but requires a large bulk while the factor B can be applied to much smaller amount. In testing these factors in the same experiment, the simple factorial scheme has to be modified in such a way that the levels of A be assigned to larger plots (units), called *Main plots* and that of B to sub-divisions of the main plots, called *Sub-plots*. Following are the examples of factorials which require large experimental units in agricultural research—

(i) In field experimentation, the factors like sowing date and irrigation cannot be applied to smaller plots while the factors like varieties and manuring can be very conveniently applied to the smaller plots.

(ii) In research on milking machine a relatively large amount of milk is required while the methods of cooling and pasteurizing require smaller amounts of milk.

(iii) In green house studies, the entire green house is used as a main plot and several treatments conducted in a green house are used as sub-plots.

In a S. P. D., the size of the main plot is considerably larger than that of a sub-plot. Hence the precision of the main plot-factor A will be much smaller than that of the sub-plot factor B and interaction AB but the average precision is the same as in the case of simple factorial experiment carried out on the same bulk of the

experimental material and with the same no. of replications. It proves that the increased precision on B and AB is obtained at the cost of the sacrifice of precision on A . Thus the *S. P. D.* can be regarded as a factorial experiment with main effect A confounded.

Advantages :

(i) The main advantage of *S. P. D.* is that two dis-similar factors as regards the necessity of the experimental material are tested in one and the same experiment.

(ii) Increased precision is obtained on the sub-plot factor and the interaction between the main plot factor and the sub-plot factor.

(iii) The inclusion of an extra factor is possible by dividing each ultimate plot into a no. of further divisions.

(iv) It saves experimental material in a no. of cases where a wide border is required between the main plots only.

Disadvantages :

(i) The main plot factor is measured with less precision.

(ii) The Statistical analysis is complex and especially when some units are missing in the data.

Applications—Keeping in view the above mentioned advantages and disadvantages, we conclude that *S. P. D.* is appropriate for the following situations—

(i) when all the factors are not of equal importance.

(ii) when one factor cannot be tested on small amounts while the other can be tested.

Randomization (when main plots arranged in a *R. B. D.*)—

For the purpose of randomization, the experimental area is divided into as many blocks as the no. of replicates and then each block is divided into as many divisions (main plots) as the no. of levels of the main plot factor. The levels of the main plot factor are randomly assigned to these divisions in each block separately. Finally, each main plot is sub-divided into as many sub-divisions (sub-plots) as the no. of levels of the sub-plot factor. The levels of the sub-plot factor are randomly allotted to the sub-plots within each main plot separately.

Statistical analysis (when main plots are arranged in a *R. B. D.*)—

First we compute the *C. F.* and the total sum of squares in the usual way and the latter is denoted by $T. S. S_1$. In the next step, we compute the sum of squares due to the main factor A , blocks and the error against which the effects of A and blocks are tested. This error

is denoted by error (a). The computations are made by arranging the data in the following tabular way—

main factor A		a_0	a_1	a_{m-1}	Totals
Blocks	B_1	$(a_0 B_1)$	$(a_1 B_1)$	$(a_{m-1} B_1)$	(B_1)
	B_2	$(a_0 B_2)$	$(a_1 B_2)$	$(a_{m-1} B_2)$	(B_2)
	\vdots	\vdots	\vdots		\vdots	\vdots
	B_r	$(a_0 B_r)$	$(a_1 B_r)$	$(a_{m-1} B_r)$	(B_r)
	Totals	(a_i)	(a_1)	(a_{m-1})	G

Where the symbol $(a_i) \rightarrow$ denotes the total yield due to i th level of A and is the sum of ' nr ' sub-plot yields. ($i=0, 1, 2, \dots, m-1$)

$(B_j) \rightarrow$ denotes the total yield of the j th block and is a sum of ' nm ' sub-plot yields. ($j=1, 2, \dots, r$)

$(a_i B_j) \rightarrow$ denotes the total yield due to i th level of A in the j th block and is a sum of ' n ' sub-plot yields.

$$\text{now, } T. S. S_2 = \sum_i \sum_j \frac{(a_i B_j)^2}{n} - C. F., \quad \text{where } C. F. = \frac{G^2}{N} \quad \text{and}$$

$$N = rmn$$

$$S. S. (A) = \sum_i \frac{(a_i)^2}{nr} - C. F.,$$

$$S. S. (\text{blocks}) = \sum_j \frac{(B_j)^2}{nm} - C. F.,$$

and $S. S.$ due to error (a) = $T. S. S_2 - S. S.$ due to ($A + \text{blocks}$)

Finally, $S. S. (AB)$ and error (b) are calculated from the following table—

		Levels of A				Totals
Levels of B	b_0	$(a_0 b_0)$	$(a_1 b_0)$	$(a_{m-1} b_0)$	(b_1)
	b_1	$(a_0 b_1)$	$(a_1 b_1)$	$(a_{m-1} b_1)$	(b_2)
	\vdots	\vdots	\vdots		\vdots	\vdots
	b_{n-1}	$(a_0 b_{n-1})$	$(a_1 b_{n-1})$	$(a_{m-1} b_{n-1})$	(b_{n-1})
	Totals	(a_0)	(a_1)	(a_{m-1})	G

$$\text{Using } S. S. (B) = \sum_K \frac{(b_K)^2}{rm} - C. F., \quad K=0, 1, \dots, n-1,$$

$$T. S. S_3 = \sum_i \sum_j \frac{(a_i b_K)^2}{r} - C. F.,$$

$$S. S. (AB) = T. S. S_3 - S. S. \text{ due to } (A + B)$$

$$\text{Error } (b) = T. S. S_1 - T. S. S_2 - T. S. S_3 + S. S. (A)$$

Now we arrive at the following *A. V. T.*—

Source of variation	<i>D. F.</i>	<i>S. S.</i>	<i>M. S. S.</i>	<i>F. cal.</i>	<i>F</i> at	
					5%	1%
Blocks	$r-1$	<i>S. S. (blocks)</i>	$S. S. (blocks)/(r-1)$	$\frac{V_{bi}}{V_{E_a}}$, if $V_{bi} > V_{E_a}$	—	—
<i>A</i>	$m-1$	<i>S. S. (A)</i>	$S. S. (A)/(m-1)$	$\frac{V_A}{V_{E_a}}$, if $V_A > V_{E_a}$	—	—
Error (<i>a</i>)	$(r-1)(m-1)$	<i>S. S. (E_a)</i>	$S. S. (E_a)/(r-1)(m-1)$	—	—	—
...
<i>B</i>	$n-1$	<i>S. S. (B)</i>	$S. S. (B)/(n-1)$	$\frac{V_B}{V_{E_b}}$, if $V_B > V_{E_b}$	—	—
<i>AB</i>	$(m-1)(n-1)$	<i>S. S. (AB)</i>	$S. S. (AB)/(m-1)(n-1)$	$\frac{V_{AB}}{V_{E_b}}$, if $V_{AB} > V_{E_b}$	—	—
Error (<i>b</i>)	$m(n-1)(r-1)$	<i>S. S. (E_b)</i>	$S. S. (E_b)/m(n-1)(r-1)$	—	—	—
Totals	$mr-1$	<i>T. S. S.</i>	—	—	—	—

S. E.—

The $S. E_s$ of the treatment-means are given below—

(1) $S. E.$ of the difference between two A means $= \sqrt{\frac{2V_{E_a}}{rn}}$,

(2) $S. E.$ of the difference between two B means $= \sqrt{\frac{2V_{E_b}}{rm}}$,

(3) $S. E.$ of the difference between two B means at the same level of $A = \sqrt{\frac{2V_{E_b}}{r}}$ and

(4) $S. E.$ of the difference between two A means at the same or different levels of $B = \sqrt{2[(n-1)V_{E_b} + V_{E_a}]/rn}$

In this case, the value of ' t ' against which the ratio $\left(\frac{\text{difference}}{S. E.}\right)$ is to be compared, is given by

$$t_{\alpha} = \frac{(m-1)V_{E_b} t_{\alpha}(\text{for error } b) + V_{E_a} t_{\alpha}(\text{for error } a)}{(m-1)V_{E_b} + V_{E_a}} \quad \text{In}$$

practice it is rarely computed.

S. P. D. (main plots arranged in R. B. D.) versus factorial experiment (arranged in R. B. D.)

S.No	S. P. D.	Factorial experiment
1	The sub-plot factor and the interaction are measured more precisely than the main plot factor but the average precision is the same as in the case of factorial experiment.	All the factors are measured with equal precision.
2	It is used when all the factors are not of equal importance.	It is used when all the factors are of equal importance.
3	The size of the plots is according to the necessity of the factors and hence the factors which require larger bulk of experimental material can be tested.	The size of the plots remains the same for all the factors and hence the factors which require relatively large experimental material cannot be tested.
4	Inclusion of an extra factor is possible without disturbing the original layout.	Inclusion of an extra factor is not possible in the pre-planned layout.

Exp. No. (1)—A Split plot experiment was laid out in four replications to study the effect of sowing dates D_1, D_2, D_3, D_4 and three depths of sowing d_1, d_2, d_3 on turmeric. The following yields in lbs. per plot was obtained—

	Rep. I	Rep. II	Rep. III	Rep. IV			
D_2	$\begin{Bmatrix} d_3 & 7.7 \\ d_1 & 5.2 \\ d_2 & 2.6 \end{Bmatrix}$	D_1	$\begin{Bmatrix} d_1 & 22.1 \\ d_2 & 19.0 \\ d_3 & 14.9 \end{Bmatrix}$	D_3	$\begin{Bmatrix} d_2 & 8.3 \\ d_1 & 5.4 \\ d_3 & 8.3 \end{Bmatrix}$	D_2	$\begin{Bmatrix} d_3 & 4.2 \\ d_1 & 7.3 \\ d_2 & 7.5 \end{Bmatrix}$
D_4	$\begin{Bmatrix} d_1 & 0.9 \\ d_2 & 2.1 \\ d_3 & 0.4 \end{Bmatrix}$	D_4	$\begin{Bmatrix} d_1 & 6.3 \\ d_3 & 3.9 \\ d_2 & 1.6 \end{Bmatrix}$	D_1	$\begin{Bmatrix} d_3 & 7.3 \\ d_1 & 7.8 \\ d_2 & 6.1 \end{Bmatrix}$	D_1	$\begin{Bmatrix} d_1 & 16.9 \\ d_3 & 10.0 \\ d_2 & 9.5 \end{Bmatrix}$
D_3	$\begin{Bmatrix} d_2 & 4.2 \\ d_1 & 1.3 \\ d_3 & 1.4 \end{Bmatrix}$	D_3	$\begin{Bmatrix} d_3 & 5.3 \\ d_1 & 6.3 \\ d_2 & 4.3 \end{Bmatrix}$	D_4	$\begin{Bmatrix} d_1 & 1.3 \\ d_2 & 1.0 \\ d_3 & 1.3 \end{Bmatrix}$	D_4	$\begin{Bmatrix} d_2 & 2.6 \\ d_1 & 4.5 \\ d_3 & 4.0 \end{Bmatrix}$
D_1	$\begin{Bmatrix} d_1 & 6.8 \\ d_2 & 6.9 \\ d_3 & 5.3 \end{Bmatrix}$	D_2	$\begin{Bmatrix} d_1 & 4.9 \\ d_2 & 4.0 \\ d_3 & 5.1 \end{Bmatrix}$	D_2	$\begin{Bmatrix} d_1 & 5.5 \\ d_3 & 2.6 \\ d_2 & 2.1 \end{Bmatrix}$	D_3	$\begin{Bmatrix} d_1 & 8.8 \\ d_2 & 6.1 \\ d_3 & 5.9 \end{Bmatrix}$

(a) Analyse the data carefully ?

(b) Give summary tables and work out standard errors and critical differences for different comparisons ?

(c) Give a statement of conclusions ?

(M. A. Patna, 1953)

Solution—

Ho : The data is homogeneous with respect to the blocks, Sowing dates and the depth of sowing and the two factors are independent.

$$C. F. = \frac{G^2}{N} = \frac{(286.8)^2}{48} = 1713.63$$

$$T. S. S_1 = \sum_i \sum_j y_{ij}^2 - C. F. = (7.7)^2 + (5.2)^2 + \dots + (5.9)^2 - 1713.63$$

$$= 2682.46 - 1713.83$$

$$= 968.83$$

To compute the S. S. for blocks (replications), factor 'D' and the error (a), we prepare the following table—

Dates Replication	D ₁	D ₂	D ₃	D ₄	Totals = B	(Totals = B) ²
I	19.0 (361.00)	15.5 (240.25)	6.9 (47.61)	3.4 (11.56)	44.8 (660.42)	2007.04
II	56.0 (3136.00)	14.0 (196.00)	15.9 (252.81)	11.8 (139.24)	97.7 (3724.05)	9545.29
III	21.2 (449.44)	10.2 (104.04)	22.0 (484.00)	3.6 (12.96)	57.0 (1050.44)	3249.00
IV	36.4 (1324.96)	19.0 (361.00)	20.8 (432.64)	11.1 (123.21)	87.3 (2241.81)	7621.29
Totals = T	132.6 (5271.40)	58.7 (901.29)	65.6 (1217.06)	29.9 (286.97)	286.8 = G (7676.72)	82254.24
(Totals = T) ²	17582.76	3445.69	4303.36	894.01	82254.24	

$$T. S. S_2 = \sum_i \sum_j \frac{y_{ij}^2}{3} - C. F.,$$

where $C. F. = \frac{G^2}{N} = \frac{82254 \cdot 24}{48} = 1713 \cdot 63$ and y_{ij} is the yield of 3 sub plots.

$$\begin{aligned} &= \frac{(19)^2 + \dots + (11 \cdot 1)^2}{3} - 1713 \cdot 63, \\ &= \frac{7676 \cdot 72}{3} - 1713 \cdot 63 = 2558 \cdot 9067 - 1713 \cdot 63 \\ &= 845 \cdot 2767 \end{aligned}$$

$$\begin{aligned} S. S. (D) &= \sum_i \frac{T_i^2}{12} - C. F. = \frac{(132 \cdot 6)^2 + \dots + (29 \cdot 9)^2}{12} - 1713 \cdot 63 \\ &= \frac{17582 \cdot 76 + \dots + 894 \cdot 01}{12} - 1713 \cdot 63 \\ &= 2185 \cdot 4850 - 1713 \cdot 63 = 471 \cdot 8550 \end{aligned}$$

$$\begin{aligned} S. S. (rep.) &= \sum_j \frac{B_j^2}{12} - C. P. = \frac{(44 \cdot 8)^2 + \dots + (87 \cdot 3)^2}{12} - 1713 \cdot 63 \\ &= \frac{2007 \cdot 04 + \dots + 7621 \cdot 29}{12} - 1713 \cdot 63 \\ &= 1868 \cdot 5516 - 1713 \cdot 63 = 154 \cdot 9216 \end{aligned}$$

$$\begin{aligned} \text{Error (a)} &= T. S. S_2 - S. S. \text{ due to (Dates + replications)} \\ &= 845 \cdot 2767 - (471 \cdot 8550 + 154 \cdot 9216) \\ &= 218 \cdot 5001 \end{aligned}$$

(iii) *S. E.* of difference between two 'd' means at the same level

$$\begin{aligned} \text{of 'D'} &= \sqrt{\frac{2V_{E_b}}{r}} \\ &= \sqrt{\frac{2 \times 3 \cdot 2581}{4}} = \sqrt{1 \cdot 62905} = 1 \cdot 2763 \end{aligned}$$

$$\begin{aligned} (C. D.) &= (S. E. \text{ of difference}) \times t \quad (24) \\ &\quad \quad \quad 5\% \quad \quad \quad .05 \\ &= 1 \cdot 2763 \times 2 \cdot 064 = 2 \cdot 634283 \approx 2 \cdot 6343 \text{ lbs/plot} \end{aligned}$$

(iv) *S. E.* of the difference between two 'D' means at the same

$$\text{or different levels of 'd'} = \sqrt{2 \left[\frac{(n-1) V}{E_b} + \frac{V_{E_a}}{E_b} \right] / m}$$

$$\begin{aligned} \therefore &= \sqrt{\frac{2[2 \times 3.2581 + 24.2778]}{4 \times 3}} \parallel \sqrt{5.1323} \\ &= 2.2654 \end{aligned}$$

$$(C. D.)_{5\%} = (S. E. \text{ of difference}) \times t_{.05}$$

$$\text{where } t_{.05} = \frac{(m-1) V_{E_b} t_{(24)} + V_{E_a} t_{(9)}}{(m-1) V_{E_b} + V_{E_a}}$$

$$= \frac{3 \times 3.2581 \times 2.064 + 24.2778 \times 2.262}{3 \times 3.2581 + 24.2778}$$

$$= \frac{75.0905}{34.0521} = 2.205$$

$$(C. D.)_{5\%} = 2.2654 \times 2.205 = 4.9952 \text{ lbs/plot.}$$

The treatment means are given in the following table—

$\begin{array}{c} D \\ d \end{array}$	D_1	D_2	D_3	D_4	Average 'd'
d_1	13.4	5.725	5.45	3.25	6.956
d_2	10.375	4.05	5.725	1.825	5.494
d_3	9.375	4.9	5.225	2.4	5.475
Average 'D'	11.05	4.892	5.467	2.492	

(i) Effect of D (Sowing Dates) :

The maximum yield has been recorded in the case of D_1 followed by D_3 and their difference is significant. The minimum yield is obtained in the case of D_4 which do not differ significantly from D_3 and D_3 . The sowing date D_1 is the best of all.

(ii) Effect of d (depths of sowing) :

The maximum yield has been recorded in the case of d_1 followed by d_2 and their difference is significant. The depths d_2 and d_3 has given almost the same average yield. Thus d_1 is the best depth of sowing.

(iii) Effect of 'dD' :

From the analysis of variance table, we see that the effect of 'Dd' is not significant. Thus the two factors are independent.

Conclusion—In order to get the maximum yield, it is recommended that the turmeric should be sown at the date ' D_1 ' and depth ' d_1 ' i. e. ' d_1D_1 ' is the optimum combination.

For calculating the S. S. due factor 'd', the interaction 'dD' and the error (b), we perform the following table—

Dates depths	D_1	D_2	D_3	D_4	Totals = d	(Totals = d) ²
d_1	53.6 (2872.96)	22.9 (524.41)	21.8 (475.24)	13.0 (169.00)	111.3 (4041.61)	12387.69
d_2	41.5 (1722.25)	16.2 (262.44)	22.9 (524.41)	7.3 (53.29)	87.9 (2562.39)	7726.41
d_3	37.5 (1406.25)	19.6 (384.16)	20.9 (436.81)	9.6 (92.16)	87.6 (2319.38)	7673.76
Totals= T	132.6 (6001.46)	58.7 (1171.01)	65.6 (1436.46)	29.9 (314.45)	286.8= G (8923.38)	

$$T. S. S_2 = \sum_i \sum_j \frac{y_{ij}^2}{4} - C. F., \text{ where } C.F. = \frac{G^2}{N} = 1713.63 \text{ and } y_{ij} \text{ is the yield of 4 plots}$$

$$= \frac{(53.6)^2 + \dots + (9.6)^2}{4} - 1713.63$$

$$= \frac{8923.38}{4} - 1713.63$$

$$= 2230.8450 - 1713.63 = 517.2150$$

$$S. S. (D) = \sum_i \frac{T_i^2}{12} - C. F. = 471.8550 \quad (\text{from 1st table})$$

$$S. S. (d) = \sum_j \frac{d_j^2}{16} - C. F. = \frac{(111.3)^2 + \dots + (87.9)^2 + (87.6)^2}{16} - 1713.63$$

$$= \frac{12387.69 + 7726.41 + 7673.76}{16} - 1713.63$$

$$= 1736.7412 - 1713.63 - 23.1112$$

$$S. S. (dD) = T. S. S_3 - S. S. \text{ due to (Dates + depths)}$$

$$= 517.2150 - (471.8550 + 23.1112) = 22.2488$$

$$\text{Error (b)} = T. S. S_1 - T. S. S_2 - T. S. S_3 + S. S. (D)$$

$$= 968.83 - 845.2767 - 517.2150 + 471.8550$$

$$= 78.1933$$

Now we arrive at the following A. V. T.—

Source of variation	D. F.	S. S.	M. S. S.	F cal.	F at	
					5%	1%
Replications	3	154.9216	51.64053	2.13	3.86	6.99
D	3	471.8550	157.2850	6.48*	"	"
Error (a)	9	218.5001	24.2778	—	—	—
			$= V_{E_a}$			
...
d	2	23.1112	11.5556	3.55*	3.40	5.61
dD	6	22.2488	3.7081	1.14	2.51	3.67
Error (b)	24	78.1933	3.2581	—	—	—
			$= V_{E_b}$			
Totals	47	968.83	—	—	—	—

(a) From the analysis of variance table, we conclude that the sowing dates (D) and the depths of sowing (d) both differ significantly at 5% level of significance and the 2 factors (sowing dates and depths of sowing) are independent.

(b) (i) S. E. of the difference between two 'D' means

$$= \sqrt{\frac{2V}{r n} E_a} = \sqrt{\frac{2 \times 24.2778}{4 \times 3}} = \sqrt{4.0463} = 2.0115$$

$$(C. D.) = (S. E. \text{ of difference}) \times t \quad (9)$$

$$5\% \qquad \qquad \qquad .05$$

$$= 2.0115 \times 2.262 = 4.550013 \approx 4.55 \text{ lbs/plot.}$$

(ii) S. E. of the difference between two 'd' means $= \sqrt{\frac{2V}{r m} E_b}$

$$= \sqrt{\frac{2 \times 3.2581}{4 \times 4}} = \sqrt{0.4073} = 0.6382$$

$$(C. D.) = (S. E. \text{ of difference}) \times t \quad (24)$$

$$5\% \qquad \qquad \qquad .05$$

$$= 0.6382 \times 2.064 = 1.3172448 \approx 1.3173 \text{ lbs/plot.}$$

EXERCISE IX

Q. (1) Describe the Split plot design and its advantages ?

The response of wheat to 4 types of composts is to be tested at 3 intensities of irrigation. State how you will layout the experiment in a *S. P. D.* and give a skelton analysis of variance of results ?

(M. Sc. Ag. Agra, 1956)

Q. (2) What is split plot design ?

An experiment is to be conducted on wheat with 5 dates of sowing and 3 varieties. Suggeste a suitable design for the experiment ? Set up the analysis of variance and indicate how you would get the sum of squares for the different components in the alalysis of variance table ?

(M. Sc. Ag. Agra, 1960)

Q. (3) (a) Describe randomized blocks and Split-plot designs in field trials and compare the two for their relative merits and demerits ?

(b) In a *S. P. D.*, there were 3 main plots with 3 varieties of paddy. The main plots were replicated 6 times. Each of the 18 main plots was Split in 4 Sub-plots on which there were 4 dates of top dressing of a fertilizer. Construct the analysis of variance table showing the sources of variation and corresponding degrees of freedom ?

(M. Sc. Ag. Agra, 1964)

Q. (4) An experiment is to be conducted for finding the effect of the three levels of irrigation on the yield of four varieties of potatoes. Suggest a suitable plan for this experiment and give the skelton of analysis of variance table ? Also indicate the method for calculating the sum of squares for the different components ?

(M. Sc. Ag. Agra, 1961)

Q. (5) An experiment was conducted for finding the effect of the 4 levels of green manuring on the yield of 3 varieties of potato and manure was applied to sub-plots while the varieties were sown in the main-plots. The following results were recorded—

Treatment :	v_1g_1	v_1g_2	v_1g_3	v_1g_4	v_2g_1	v_2g_2	v_2g_3	v_2g_4
Total yield :	429	538	665	711	480	591	688	749
of 6 plots					v_3g_1	v_3g_2	v_3g_3	v_3g_4
(in seers)					520	651	703	761

The Experimental Designs

16

ex

pr

ce

re

A

f

.

Source	D. F.	S. S
Replications	—	15875.278
Varieties	—	—
Error (a)	—	6013.30
...
Manuring	—	—
Interaction	—	—
Error (b)	—	—
Totals	71	51985.944

Carryout the further analysis and state your conclusions ?

Ans: Blocks and manuring differ significantly.

CHAPTER X

Switch-over Trials

Or

Cross Over Designs

The experimental Designs discussed so far pertain to situations in which the treatment applied to a particular experimental unit remains the same for the whole duration (period) of the experiment. The designs discussed in the present chapter pertain to situations in which the treatment applied to a particular experimental unit does not remain the same for the whole duration of the experiment. But the whole duration is divided into as many fractional periods as the no. of treatments and the treatments are assigned randomly to these fractional periods. *Since the treatments are switched in sequence over several fractional periods, hence the design is called switch over design.*

Applications : This design is used in dairy husbandry, biological psychological, and marketing research.

Advantages : (1) This design is very useful for the situations where the effect of the treatment varies with the time and the whole period can be divided into different fractional periods according to these effects.

(2) This design estimates the treatment effect over a short period of time and controls the fluctuations due to the time.

Randomization : Since all the treatments are applied to each experimental unit, hence each unit is considered as a full replicate. For the purpose of randomization, the whole duration of the experiment is divided into as many fractional-periods as the no. of treatments and the treatments are randomly allotted to these fractional periods within each replicate with the restriction that each treatment is to be given an equal no. of times in each fractional period. It requires that the no. of replicates must be an exact multiple of the no. of treatments. The treatments are arranged in a R. B. D. or L. S. D.

Following is the plan for two treatments (A and B) and 8 replicates in a R. B. D.—

Replicate \ Period	1	2	3	4	5	6	7	8
I	B	A	B	A	A	B	B	A
II	A	B	A	B	B	A	A	B

If the experiment has to be conducted in a L. S. D, the plan will be of the following form—

Square I

A	B
B	A

Square II

B	A
A	B

Square III

B	A
A	B

Square IV

A	B
B	A

Statistical Analysis : For a switch-over design arranged in a R B. D. with 'n' replicates and 'k' treatments, the break down of the d f. is as follows—

Source of variation	D. F.
Replicate	$n-1$
Period	$k-1$
Treatment	$k-1$
Error	$(n-2)(k-1)$
Totals	$nk-1$

The computations of the sum of squares are shown in the following example—

Exp. (1) : Two rations A and B were administered to 8 dairy-cows. Each cow received ration A and B in period I (first half of the lactation period) and period II (second half of the lactation period). The rations A and B were allotted to the two periods at random with the restriction that half of the cows received the ration A and the other half the ration B in each period. The experimental design and the milk yield (in seers) are given in the following table—

Cows \ Periods	1	2	3	4	5	6	7	8
I	B 25	A 23	B 15	A 20	B 15	B 15	A 14	A 20
II	A 15	B 11	A 15	B 15	A 15	A 12	B 6	B 14

Analyse the data and interpret the results ?

H_0 : The data is homogeneous.

The Experimental Designs

184

$$C. F. = \frac{G^2}{N} = \frac{100}{16} = 6.25$$

$$T. S. S. = \sum_i \sum_j y_{ij}^2 - C. F. = 322 - 6.25 = 315.75$$

$$S. S. \text{ due to cows} = \sum_j \frac{C_j^2}{2} - C. F.$$

$$= \frac{100 + 16 + \dots + 16}{2} - 6.25$$

$$= 133 - 6.25 = 126.75$$

$$S. S. \text{ due to periods} = \sum_i \frac{P_i^2}{8} - C. F. = \frac{729 + 289}{8} - 6.25$$

$$= 127.25 - 6.25 = 121.00$$

$$S. S. \text{ due to Rations} = \frac{T_A^2 + T_B^2}{8} - C. F. = \frac{196 + 16}{8} - 6.25$$

$$= 26.5 - 6.25 = 20.25$$

$$S. S. \text{ due to error} = T. S. S. - S. S. \text{ due to (cows + periods + rations)}$$

$$= 315.75 - (126.75 + 121.00 + 20.25)$$

$$= 47.75$$

Now we arrive at the following A. V. T.—

Source of variation	D. F.	S. T.	M. S. S.	F cal.	F at	
					5%	1%
Cows	7	126.75	18.1071	2.275	4.21	8.26
Periods	1	121.00	121.00	15.204**	5.99	13.74
Rations	1	20.25	20.25	2.545	"	"
Error	6	47.75	0.9583	—	—	—
Totals	15	315.75	—	—	—	—

Inference : The two rations A and B do not differ significantly while the two periods differ significantly at 1% level.

If necessary, the S. E. of the difference between the two treatment means is given by S. E. of difference = $\sqrt{\frac{2V_E}{r}}$

EXERCISE X

Q. (1) : What is switch over trial ?

Give in detail the plan for feeding trial conduct with the object of testing the effect of three different rations on the milk yield of Harijina cows. Indicate briefly the method of analysis of the data ?

(M. Sc. Ag. Agra, 1961)

Hint : Let the three rations be A, B and C. The whole lactation period will be divided into 3 fractional periods I, II and III. In this case, the no. of replications may be 6 or 9 or 12 etc. the exact multiple of 3. Suppose we start with 6 replications, then the plan will be of the followin formg—

Cows / \ Periods	1	2	3	4	5	6
	I	II	III	I	II	III
I	A	A	B	C	C	B
II	C	B	C	B	A	A
III	B	C	A	A	B	C

The break down of d. f. will be as follows—

Source of variation	D. F.
-----	----
Cows	5
Periods	2
Rations	2
Error	8
-----	----
Tatals	17
-----	----

CHAPTER XI

Progeny Row Trials

And

Compact Family Block Designs

The selection of plants for further propagation is of paramount importance in plant breeding work. In old days, the method of mass selection was in use for the selection of such plants. The method consists in choosing from the material under selection, a no. of plants that appear to be superior in respect of the character or characters under selection, bulking the seed from these selected plants, raising from the seed a next generation and again choosing the superior plants from this generation and repeating the same operation as before. The main drawback of this method was that in each generation the selection is subject to environmental or non-genetic variability present in the field. Due to this drawback, this method was inefficient and defective.

After the development of the basic principles of experimental designs, the necessity of the objective testing by the application of randomization and replication was realized. The replicated experiments require a comparatively large amount of seed and the experimental-material. But the amount of seed produced by the selected plants remains small which presented a difficulty in the application of randomization and replication. Another difficulty in conducting the replicated experiments was that the genetic variation (due to the heterogeneity of the plant breeding material) enters the variation due to error. These two difficulties are successfully overcome by adopting the method of '*Progeny Row Trials*'. This method is statistically sound being based on the principles of randomization and replication and with extremely small plot-size. The present method consists in

sowing the seeds of different selected plants in separate rows and selecting that plant which seems to be superior on the average. In this method, the average performances of progenies are taken into account while the method of *mass selection* depends upon the performance of the individual plants. It is a well known fact that the mean is subject to a fraction of the environmental variation to which the individual plants are subjected. Hence this method is more reliable.

The progeny row trial is a randomized block design (R. B. D.) in which each plot consists of 3 rows. All the seeds of a plot belong to the same parent plant (progeny). The middle row is called the *progeny or experimental row* and the two side rows are called *the guard rows*. The guard rows avoid the border-effect on the experimental row.

Statistical Analysis—Suppose we have ' p ' progenies and ' r ' replications. Then the break down of the *d. f.* will be as follows—

Source of variation	<i>D. F.</i>
Replication	$(r-1)$
Progeny	$(p-1)$
Error	$(r-1)(p-1)$
Total	$rp-1$

The computations of the sum of squares due to different sources and test of significance are made exactly in the same way as in the case of R. B. D.

It *F*-test indicates that the different progenies differ significantly, they are arranged in the descending order of their magnitudes. The progeny with maximum mean is selected for further propagation provided it differs significantly from the remaining progenies. In the case, when the maximum mean value does not differ significantly

from one or more mean values, the progeny of mean with greatest plant error (the variance between the plants of the same progeny) is selected from those which do not differ significantly

Compact Family Block Design—In Progeny row trial, all the progenies to be compared belong to the same family. However, if the progenies to be tried belong to the different families then compact family block design which is analogous to the Split Plot Design (S. P. D.) is used to compare—

- (1) Different families,
- (2) Progenies belonging to the same family, and
- (3) Progenies belonging to the different families.

Randomization—In this design, the families are randomized in the main plots and the progenies of the same family are randomly allotted to the progeny sub-plots within the main plots.

Statistical Analysis—Suppose we have ' f ' (F_1, F_2, \dots, F_f) families, ' p ' progenies in each family and ' r ' replications. Then the above mentioned 3 comparisons are made in the following way—

(1) **Comparison between families**—To have a comparison between the families, the main-plot data are analysed. The break down of *d. f.* is as follows—

Source of variation	<i>D. F.</i>
Replications	$r-1$
Families	$f-1$
Error	$(r-1)(f-1)$
Totals	$rf-1$

The computations of sum of squares and the test of significance are made exactly in the same way as the main plot data is analysed in the case of Split Plot Design.

The *S. E.* of the difference between the 2 family means

$$= \sqrt{\frac{2 \times \text{Error variance between families}}{rp}}$$

(2) **Comparison between progenies within families**—For this purpose, the various families are analysed separately and tested against the error obtained from the data of that family. The results are arranged in the following tabular form —

Source of variation	D. F.	F_1				F_f			F at	
		$S. S$	$M.S.$	$F. cal$...	$S. S.$	$M.S$	$F. cal$	5%	1%
Replications	$(r-1)$
Progenies	$(p-1)$
Error	$(r-1)(p-1)$	—	—	—	—
Totals	$(rp-1)$	—	—	—	—	—	—	—	—	—

The *S. E.* of the difference between the two progeny means within the same family

$$= \sqrt{\frac{2 \times \text{Error variance between the progenies within the same family}}{rf}}$$

(3) **Comparison between progenies belonging to different families**—If the within family error variances are homogeneous then the different families and their progenies are compared with the help of the following pooled analysis of variance table—

Source of Variation	D. F.	$S. S.$	$M. S. S.$	$F. cal.$	F at	
					5%	1%
Replications	$(r-1)$
Families	$(f-1)$
Error (a)	$(r-1)(f-1)$...	V_{E_a}	—	—	—
Progenies within families	$f(p-1)$
Error (b)	$f(p-1)(r-1)$...	V_{E_b}	—	—	—
Totals	$frp-1$	$T.S.S.$	—	—	—	—

$S.S.$ due to progenies within families and error (b) are computed by adding the corresponding sum of squares obtained for different families.

• •

The $S. E.$ of the the difference between the two progeny means belonging to different families

$$= \sqrt{\frac{2}{r} \left\{ \frac{V_{E_a} + (p-1) V_{E_b}}{p} \right\}}$$

Advantages of Compact family Block Designs—

(1) The main advantage of this design lies in the fact that the progenies of a family are shown side by side in a family main plot. Hence the progenies of a family experience the same type of environment, and variation within families is minimum.

(2) When the analysis of variance indicates that certain families are inferior to others, further analysis of these families is not essential. This saves a considerable amount of time and labour involved in the analysis when a large number of families is tried.

EXERCISE XI

Q. (1) Describe the compact family block design for plant breeding trials and discuss its advantages. Give the skeleton of analysis of variance of such a trial ?

(M. Sc. Ag. Agra, 1958)

Q. (2) Write short notes on—

(i) Replicated progeny row trials,

(M. Sc. Ag. Agra, (1956)

(ii) Compact family block designs,

(M. Sc. Ag. Agra, 1960, 64)

CHATER XII

Rotational Experiments

Certain agronomic experiments are conducted on the same experimental site for a number of years to compare the several crop rotations or sequence of agronomic practices. These experiments are called *the Rotational Experiments*.

For illustration of the procedure of randomization and statistical analysis, let the different rotations to be compared be—

R_1 (Lubia—wheat)

R_2 (Moong—wheat)

R_3 (Dhancha for green manuring—Wheat)

R_v (no cropping since April to October—wheat) and the experiment be repeated for ' n ' years in ' r ' blocks.

Randomization—The experimental site will be divided into ' r ' blocks and each block will be divided into ' v ' plots. Then ' v ' rotations will be randomized in each block separately. The experiment will continue for ' n ' years.

Statistical Analysis—The data for each year will be analysed separately. The breakdown of the degrees of freedom for the data of ' n ' years is as follows—

Source of variation	D. F.
Blocks	$(r-1)$
Rotations	$(v-1)$
Error	$(r-1)(v-1)$
Total	$rv-1$

There will be ' n ' analysis of variance tables, one for each year. If the error variances are homogeneous then the 'pooled data can be analysed. The breakdown of the d. f. will be as follows—

Source of variation	D. F.
Blocks	$n(r-1)$
Years	$n-1$
Rotations	$v-1$
Interaction ($R \times Y$)	$(v-1)(n-1)$
Error	$n(r-1)(v-1)$
Total	$nrv-1$

The sum of squares due to blocks and error will be obtained by adding the corresponding *S. S.* obtained from the individual analysis of variance tables and the *S. S.* due to rotations, years and their interaction will be obtained by arranging the data in the following $v \times n$ table—

Rotation \ Year	Y_1	Y_2	Y_n	Totals
R_1	(R_1Y_1)	(R_1Y_2)	(R_1Y_n)	...
R_2	(R_2Y_1)	(R_2Y_2)	(R_2Y_n)	...
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
R_v	(R_vY_1)	(R_vY_2)	(R_vY_n)	...
Totals	\vdots	G

• The rest procedure is the same as in the case of $(v \times n)$ factorial experiments.

If the rotations differ significantly, then the *S. E.* and *C. D.* will be computed for selecting the best rotation.

The *S. E.* of the difference between the two rotation means

$$= \sqrt{\frac{2V_E}{rn}}$$

Advantages—(1) These experiments are conducted to select the best rotation for a given locality.

(2) The rotational experiments are also used to compare the agronomic practices on a fixed rotation of crops.

EXERCISE XII

Q. (1) Describe in short “Simple Rotational Experiments” and give the procedure of randomization and statistical analysis by considering a suitable example ? Also mention the object of these experiments ?

Part III

Official Agricultural Statistics

Chapter I

Official Agricultural Statistics

Definition : Official Agricultural Statistics is defined as the aggregate of quantitative information bearing on the different fields of agriculture and its economy.

Classification : It covers a very wide field which can be classified as follows :—

- (i) Land Utilization Statistics,
- (ii) Agricultural Production Statistics including Live-Stock and Fisheries,
- (iii) Agricultural Price and Wage Statistics,
- and (iv) Ancillary Agricultural Statistics.

Importance : In a country like India which is predominantly an agricultural country, the collection of agricultural statistics is of paramount importance. Since these statistics are directly connected with the rural economy and formulation of the food policy which is of utmost importance to the whole country. Most of the original agricultural statistics of our country is collected in connection with the area and yield of different crops and crop-forecasting. They are very helpful to the government in the formulation of agricultural development plans and food policies, measurement of the effect of the past development policies and collection of land revenue. They are also useful to the traders and general public as they help in stabilizing the prices of agricultural commodities. The collection of live stock and fisheries statistics is of equal importance as they solve a good deal of food problem. Keeping all these facts in view, we conclude that the collection of statistics regarding the crop-acreage and production, live stock and their products and fisheries is of primary importance.

Agency for Collection : In India, the importance of agricultural statistics has all through been realized. The statistics pertaining

to the agriculture are some of the earlier statistics that we have. Kautilya's Arthashastra as also Aine Akbari provide a lot of informations regarding the population, acreage of crops and prices of commodities etc. The modern history dates back to the year 1875 when the department of Agriculture and commerce was set up in Uttar Pradesh and later on in 1881 similar departments were opened in other provinces as a result of the recommendations of the Famine Commission of 1880. After the First World War, certain improvements were made in the system of collection of agricultural statistics. The coming of independence led further demand for the co-ordination of the statistical material relating to agriculture. At present, all work relating to the collection, compilation and publication of agricultural statistics is carried out by the directorate of Economics and Statistics in the Central Ministry of Food and Agriculture. The following are some of the important regular publications of this directorate—

- (i) Abstract of Agricultural Statistics of India,
- (ii) Indian Agricultural Statistics (Vol. I and II),
- (iii) Estimates of area and production of principal crops in India (Vol. I and II),
- (iv) Agricultural prices in India,
- (v) Agricultural wages in India,
- (vi) Bulletin of Food Statistics,
- and (vii) Indian agriculture in brief. .

The research work for the improved methodology in the field of agricultural statistics is carried out by the statistics branch of Indian Council of Agricultural Research (I. C. A. R.). It works in close co-operation with the directorate.

With a view to collect comprehensive information relating to all sections of national economy the directorate of National Sample Survey (N.S.S.) under the department of economic affairs ministry of finance was established in June, 1950. Since in 1953 the N.S.S. has taken over the work of large scale sample survey in the field of agricultural statistics which was previously conducted by I.C.A.R. In U. P.; the work relating to agricultural statistics is carried out by the directorate of agriculture. The districtwise information relating to land utilization, area and yield of principal crops and some other useful informations are published annually in the bulletin of agricultural statistics for Uttar Pradesh.

EXERCISE No. (1)

Q. No. (1) Write a short note on the importance of agricultural statistics for any country ? (*M. Sc. Ag. Agra 1965*)

Q. No. (2) Name six important official publications relating to agricultural statistics in India ?

Q. No. (3) What are the agencies for the collection of agricultural statistics ?

Chapter II

Land Utilization Statistics

Introduction : The statistics relating to land-utilization are being collected since 1884. In U.P., districtwise detailed statistics of land utilization, area and production of crops, live-stock, agricultural prices and other useful informations with a description on the weather and crop conditions are published in “*season and crop report of U. P.*” It is an annual publication, published by the “*Board of Revenue U. P.*”

Method of collection in temporarily settled areas : In the temporarily settled areas like U. P., detailed crop records on statistics are kept by the village accountant (called *Lekhpal* in U. P., *telathi* in Maharashtra, *karnam* in South and *karamchari* in Bihar) and supervised by his immediate officer *kanungo*. For a better supervision and random checking, some state governments (U. P. is one of them) have appointed *District Statistical Officers* (D. St. O.). The statistics thus obtained are fairly reliable.

Method of collection in permanently settled areas : In permanently settled areas like Bihar and Bengal, no detailed crop records are required to be kept for revenue purpose and this work is entrusted to the *police chokidars* of the respective villages or village head men. These statistics are merely the guess work and hence they are not reliable and accurate.

Classification : Land utilization statistics are classified under the following heads—

- (1) *total area and classification of area,*
- (2) *irrigation area,*
- and (3) *cropwise area.*

(1) (a) Total area : The geographical area is furnished by village returns prepared by *Lekhpal*. It is exclusive of corporation,

municipal and town areas. These areas are based on *cadastral survey* carried out by the state government.

(b) Classification of area : In 1949-50, the area was classified into the following nine classes—

(i) Forest : This class includes all forested areas on the land or administered as forest under any legal enactment dealing with forests whether state owned or private.

(ii) Barren and unculturable land : It stands for all barren and unculturable land like mountains, usar land etc. which cannot be brought under cultivation.

(iii) Land put to non-agricultural uses : It gives the area covered with water, sites, roads, buildings, cremation ground etc. and all other lands put to the uses other than agricultural.

(iv) Cultivable waste : This category covers the land available for cultivation but not taken up for cultivation or abandoned after a few years for one reason or the other.

(v) Permanent pastures and other grazing lands : This heading includes all the grazing lands whether or not they are permanently pastures and meadows.

(vi) Land under miscellaneous trees, groves, not included in area sown : It has all culturable land which is not included in the area sown but it is put to some agricultural use. Tanel under groves, forests of timber and fuel trees, shrubs, bushes etc. which are not included under orchard are kept in this category.

(vii) Current fallows : This category comprises cropped areas, which are kept fallow during the current year.

(viii) Other fallow lands : It includes all the lands which are taken up for cultivation but are temporarily out of cultivation for a period of not more than five years.

(ix) Net area sown : It has the net area sown with crops and orchards.

(2) Irrigated area : The irrigated area is classified according to—

- (a) *source of irrigation,*
- and (b) *crops irrigated.*

The data under category (a) stands for the net area irrigated while that under (b) represents the gross irrigated area.

(3) **Cropwise area :** Fairly detailed informations are collected relating to acreage under important crops. The total cropped area is divided into—

- (a) *area under food crops,*
- and (b) *area under non-food crops.*

The former group is further sub-divided into cereals, pulses, sugarcane, coneliments and spices, fruits and vegetables including root-crops and others. The later is further classified into oil seeds, fibres, dyes and tanning materials, drugs and narcotics, fodder crops, green manure crops and others.

Causes of inaccuracies : Following factors are responsible for inaccuracies in the collection of area statistics—

- (i) lack of training and efficiency of the Lekhpal,
- (ii) heavy work load and low salary of the Lekhpal,
- (iii) lack of supervision and random checking,
- (iv) errors committed in the compilation of figures,
- (v) use of mixed crops, (no definite formula can be given for the accurate estimation of the acreage under mixed crops as the composition of the crops changes very often).
- (vi) use of fixed ridges; (the field ridges are also included in the estimate of area. These are neither sown nor cropped.)
- (vii) lack of definition of acreage under crop; (it is not certain whether acreage under crop means the area sown or the area successfully cropped.)
- (viii) method of collection adopted in permanently settled areas.

Suggestions for the improvement of the statistical data : Following measures have been suggested for improvement of the statistical data.

- (i) The steps should be taken to train the lekhpals in the technique of data-collection.

- (ii) The jurisdiction of the lekhpal should be reduced and he must be better paid.
- (iii) The supervisors should pay the surprise visits to verify the accuracy of their subordinate members.
- (iv) A certain % of the area should be considered as occupied by the ridges and the total acreage figures should be accordingly adjusted.
- (v) For the accurate estimation of acreage under mixed crops, the *method of crop cutting experiments* should be followed.
- (vi) In permanently settled areas, *random sampling method* should be adopted to obtain the reliable statistics.

EXERCISE No. 2

Q. No. (1) Write an essay on *Land-utilization Statistics* ?

Q. No. (2) Name the classifications of *Land-utilization Statistics* and explain the defects of collecting the area statistics ? Also supply some suggestions for improvement in the collection of the area statistics ?

Chapter III

Method of Estimating Crop-Yield

There are two methods of estimating the crop yield—

(i) **Annawari Estimation :** According to this method, the yield of a crop in a year for a district is estimated by the following formula—

District yield (yd) = $\frac{1}{16}$ [Area \times Normal yield \times Anna condition of the crop]

Area : The acreage under a crop is furnished by the village-papers prepared by leklpal and supervised by the supervisor kanungo. The area statistics thus furnished are fairly accurate in this state (U. P.).

Normal yield : It is the *average yield on the average soil in a year of average character as deduced from a consideration of the informations obtained on the experiments made during the year under review.* The state department of agriculture is responsible for the estimation of normal yield which is obtained on the basis of *crop-cutting experiment.* The method consists in selecting some average plots and sowing and harvesting the crop on these plots.

Anna-condition of the crop : Taking the normal yield as 16 annas, the condition of a crop in a particular year can be described in relation to the normal yield in terms of annas. For instance, if a crop seems to be $\frac{3}{4}$ of the normal crop then its yield is taken as 12 annas. The *anna-condition* of the crop is based merely on eye inspection of the crop reporter.

This method has been subjected to the following criticisms—

- (1) The selection of the plots and estimation of anna-condition are not based on any objective-method but fully depend on the discretion of the experimenter.

- (2) The standard errors of the estimates cannot be measured.
- (3) The districts selected for crop-cutting experiments continue to be those which were approved some 50 years back.

(ii) **Random Sampling Method** : The method which is used these days, was suggested by *P. V. Sukhatme* in *Agricultural situation in India, November, 1954*.

This method contains the following main steps—

- (1) The district is stratified into administrative sub-divisions (Pargana or Block etc.).
- (2) We select a number of villages in proportion to the area such that crops are randomly selected from each sub-division.
- (3) Two fields are selected at random from each of the selected villages.
- (4) A plot of desired size and shape is located in each of the selected fields.

On the basis of the yields of these plots, the estimates of the yield/acre and its standard error are computed. If A_i stands for the area of i^{th} sub-division and y_i (estimated from the sample data) for the average yield in the i^{th} sub-division, then the average district-yield (y_d) is given by the formula—

$$y_d = \frac{\sum A_i y_i}{\sum A_i}$$

Further, if S^2_v denotes the variance between the villages (in A. V. T.) and m_i denotes the number of villages in the i^{th} sub-division then

$$V(y_d) = \frac{\sum \frac{A_i^2}{m_i} \times S^2_v}{2(\sum A_i)^2}$$

This method is more scientific and reliable as compared to the traditional (*Annawat*) method due to the following reasons—

- (1) The selection of the villages, the fields, within the selected villages and the location of the plots within

- the selected fields is done at random and does not depend on the discretion (at any stage of selection) of the investigator. Thus the estimates obtained from the random sampling experiments are found as unbiased.
- (2) The estimates are based on the *modern statistical technique of sampling* and not on the guess of the investigator as in the previous method. Therefore, the results are very near to the true values.
 - (3) The standard errors of the estimated values can be measured.

Yield-Estimation

Size and Shape of the plot : The following table shows the size and shape of the plots which are adopted in U. P. for different crops—

<i>Name of crop</i>	<i>Shape</i>	<i>Size</i>
(i) Sugarcane	square	33' × 33'
(ii) Jute	square	16.5' × 16.5'
(iii) Cotton	rectangle	66' × 33'
(iv) Others	Equilateral triangle	side 33'

Location of the plot within the field : To locate the plot within the field, the South-West (S. W.) corner of the field is chosen as the starting point and a peg is fixed there. The length and the width of the field are measured in terms of the steps from the starting point. Then two random numbers are selected such that 1st of them does not exceed the difference of the number of steps along the length and that of 13 and second does not exceed the number obtained by subtracting 11 from the number of steps taken along the width. We may assume x and y two such numbers respectively along the length and the width.

Now starting from the peg at the S. W. Corner, x steps are measured along the length and y steps along the width (perpendicular to the length). A second peg is also fixed at this newly arrived (x, y) point. If a triangular plot is to be located within the field then one vertex of the equilateral triangular chain is kept fixed at the point (x, y) and the chain is stretched in such a way that one of its sides remains parallel to the length and away

from the starting point. A third peg is fixed at the second vertex of the chain and finally the third vertex of the chain is continued to be moving along the direction of the width away from the starting point till the chain is fully stretched and then a fourth peg is fixed at the third vertex of the chain. The equilateral triangle with vertices marked by the second, third and fourth pegs will be a desired plot.

Crop Fore-cast : For the purpose of administration and policy formulation, the knowledge of the acreages under important crops and their expected yields is very essential. All these informations with a description of the general conditions of the crops are published from time to time during the growth-period of the crops in the form of *bulletins*, one for each crop. *These bulletins are known as the crop fore-casts.*

Three bulletins are issued for most of the crops. The first bulletin is issued after *one month of sowing* the crop which contains the informations regarding the area sown and the conditions of germination. The second bulletin is issued *after 3 months of sowing* which gives an idea of the crop-condition and the anticipated yield of the crop. The third bulletin is issued (published) *about a-month before the harvesting* of the crop which gives the idea of the crop-yield to be harvested.

As a matter of fact, the number of fore-casts depends upon the importance of the crop. For instance, the number of fore-cast is one only in the case of *Castor, Ginger and Groundnut*, two for *Jowar, Bajra, Maize* and for *Kharif and Rabi pulses*. The number of fore-casts is 5 for *Rice and Wheat* and it is 6 for *Sugar-cane*.

The important crops for which the crop fore-casts are not made so far, are *Tobacco and Potato* but the plan for their fore-cast is under consideration.

Live stock : At present a quinquennial live stock census including details on agricultural implements and machinery of all types is held in every part of the country. The collected informations are published in *India Live Stock Census*. The last census was conducted in 1961 on an improved basis. Uniform definitions were adopted by all the states, and live stock census officers were appointed in each state. The provision was also made

for rationalized supervision and for training to enumerators and supervisors.

In 1961, live stock was classified into two broad groups :

(i) Bovine and (ii) Other.

The *bovines* were classified as cattle and buffaloes and further classified according to the sex and then age. *Other live stock* includes sheep, goats, horses and ponies, mules, donkeys, camels and pigs. The sheep, goats, horses and ponies are classified according to the age and sex while the donkeys and pigs are classified according to age only.

In the same year *i.e.* in 1961, the poultry was classified as :

(i) Fowls (ii) Ducks and (iii) Others.

The *Fowls* include hens, cocks and chickens and *Ducks* include ducks, ducklings and drakes.

Fisheries Statistics

The *Fisheries Statistics* are very inadequate and highly unorganised. The available data can be classified in the following four classes :

(1) *Data available in market-reports.*

(2) *Data available with Fisheries Research Institutes and stations* and they are namely :

- (a) Central Inland Fisheries Research Station at Calcutta,
- (b) Central Marine Fisheries Research Station at Mandapam Camp,
- (c) Deep Sea Fisheries Station at Bombay,
- (d) Offshore Stations at Tuticorin, Cochin and Vishakhapatnam,
- (e) Central Fisheries Technological Research Station at Cochin and
- (f) Fisheries Extension Units.

(3) *Data available with Fisheries Development Adviser and in State Gazettes.*

(4) *Data about consumption of fish collected by N. S. S.*

Our *Five Year Plan* have the provision for the development of fisheries in India. The F. A. O. is also helping in the development-schemes. It is hoped that in future, the position of *Fisheries Statistics* will be much improved in India.

Shortcomings and Improvements in Agricultural Statistics

The main defects of *Indian Agricultural Statistics* may be classified under the following heads :

- ... (i) Gaps in coverage ;
- (ii) Lack of uniformity in definitions ;
- (iii) Defects of primary reporting agency ;
- (iv) Defects of tabulation and processing ;
- (v) Defects of inspection, supervision and checking ;
- (vi) Defects of planning and co-ordination and
- (vii) Delay in publication.

(1) **Gaps in Coverage :** *Agricultural Statistics* are not available for the whole geographical area of India. For several million acres in *Rajasthan, Gujrat, Assam* and *Kashmir*, figures are not available. Some of the areas are also not cadastrally surveyed and hence agricultural statistics in these areas are not properly organised. Further there are certain areas where reporting agencies do not exist and as such estimates for a number of crops are not available. The agricultural statistics of India can be said to be complete when estimates of acreage under different crops and land-use classes become known in respect of the whole geographical area of the country.

During the last few years steps are taken to remedy the above defects. The *reporting area* has been increased ; *reporting agencies* have been set up in the area where they did not exist ; and regular estimates are being published for several minor crops.

(ii) **Lack of uniformity in definitions :** The methods of obtaining the *area statistics* are different in *temporary areas* and *permanently settled areas*. The position has improved to a considerable extent in *Bihar* and *Bengal*, and steps are being taken to improve the position in *Orissa*.

The definitions of the different land-use classes depend on *local customs* and usages, and were not, therefore uniform. For example, the definition of *current fallows* varied in different states. Now the Government of India have increased the number of land-use classes and have also laid down the uniform definitions for all the states in India.

The methods of yield estimation are also not uniform throughout the country, as in some states the *Annawari system*

is used while in Punjab the *Method of Direct Estimation* is followed. For some of the crops the *Random Sampling Technique* is used in some states. But the attempts are being made to follow one and the same common method of estimation in all the states of the country.

(iii) Defects in tabulation and processing :

A good deal of the data collected is rendered useless, as no processing is done. For instance in Punjab, Delhi and Madhya Pradesh (M. P.) the informations regarding the transfers of agricultural properties are collected for each village by the lekhpals; but these are not consolidated for the whole state. The similar is the case with other states,

(iv) Defects of the primary reporting agencies :

The defects in this connection have already been discussed both for the *temporarily settled areas* and the *permanently settled areas*. The *comparatively area statistics* are more reliable in the temporarily settled areas. Even if the area recorded by lekhpals is correct, his classification for different crops may not be correct. One of the chief causes for the defects in the reporting of the primary agency is the heavy work-load of the lekhpals. State Governments are now taking steps to reduce the load of the lekhpals. Sometimes the bias of the primary reporting agency also accounts for the defect in the figures.

(v) Lack of Supervision : The work of the lekhpals is inspected by his immediate officers, Kanungoes and Tehsildars. But on account of their heavy administrative duties they are not always able to devote the personal attention to such inspection works. This indifference induces the lekhpals to be negligent in their duties. State Governments are now insisting on better supervision and checking of the work of the primary reporting agency. Sometimes the villages for inspection are selected by random sampling methods. The work of a Kanungo is also being reduced to make their supervision and checking better.

(vi) Lack of Co-ordination : Sometimes the data on the same subject are collected by two or more agencies. For example, in some states the *department of agriculture* and that of *civil supplies* obtain the estimates of food production independently of each other. The former department is concerned with the increase of yield of the crops, while the later with procurement and equitable distribution of food. Often, the data collected are different and hence there is a great need of co-ordination in such cases,

(vii) **Delay in Publication:** The data are first collected in the books (registers) of lekhpals and are consolidated first for each village, then for each tehsil, then for each district, then for state as a whole ; and each state then sends its consolidated returns to the *Directorate of Economics & Statistics* at the centre for consolidation and publication on all India basis. In this system if the delay occurs at village or tehsil level, it causes delay in publication at all India level. Hence it is often suggested that closer insistence on punctuality at every stage of this system should be made.

EXERCISE No: 3

Q. No. (1) : Write short notes on the following :—

(a) Improvement of agricultural statistics in India.

(*M. Sc. Ag. Agra, 1961*)

(b) Crop-cutting experiments in U. P.

(*M. Sc. Ag. Agra 1962*)

(c) Live stock census;

(*M. Sc. Ag. Agra, 1962, 1964*)

Q. No. (2)

(a) Discuss the defects of *annawari* estimation of crop production ?

(b) In a random crop-cutting survey for estimation of yield of wheat crop describe step by step the procedure for locating a 33' equilateral triangle for sample cut in a wheat field. The yields from 2 such sample cuts are 6 seers 1 ch. and 11 seers respectively. Express these yields on a per acre basis ?

(*M. Sc. Ag. Agra 1963*)

Q. No. (3) : Write short notes on the following :—

(a) Season and crop report of U. P. (*M. Sc. Ag. Agra 1963*)

(b) Importance of agricultural statistics for any country,

(*M. Sc. Ag. Agra 1965*)

(c) Recent improvements in the sphere of—

(i) land utilization statistics,

and (ii) yield and production statistics.

(*M. Sc. Ag. Agra 1965*)

(d) Sample cut of 33' equilateral triangle.

Q. No. (4) : What is '*Season and crop Report of U. P.*' ? Who publishes it, and what is the frequency of its publication ? List the types of information contained in this publication and discuss the utility of the same in a planned economy ?

(*M. Sc. Ag. Agra 1964*)

Q. No. (5) : The shape of a wheat field is nearly rectangular and the measures of its length and breadth from the south-west are 212 steps and 98 steps respectively. Explain clearly with the help of a diagram how you would locate a 33' equilateral triangle for crop cutting experiment in this field ? (*M. Sc. Ag. Agra 1964*)

Q. No. (6) : Write an essay on agricultural statistics.

Appendix I & II
Tables & A. U. Papers

